CDL in SCIENZE RIABILITATIVE DELLE PROFESSIONI SANITARIE

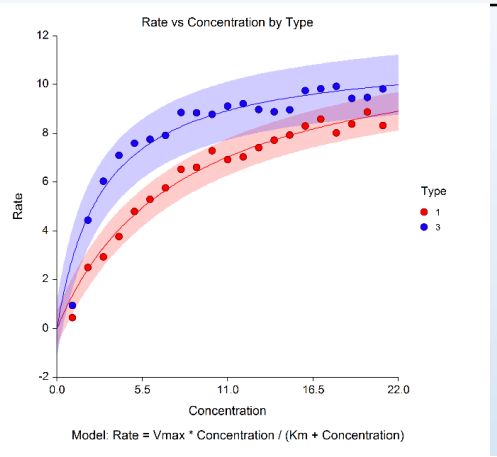
STATISTICA PER LA RICERCA

gbarbati@units.it

A.A. 2025-26

- Concetti generali della Statistica Descrittiva

- Richiami di rappresentazione grafiche dei dati



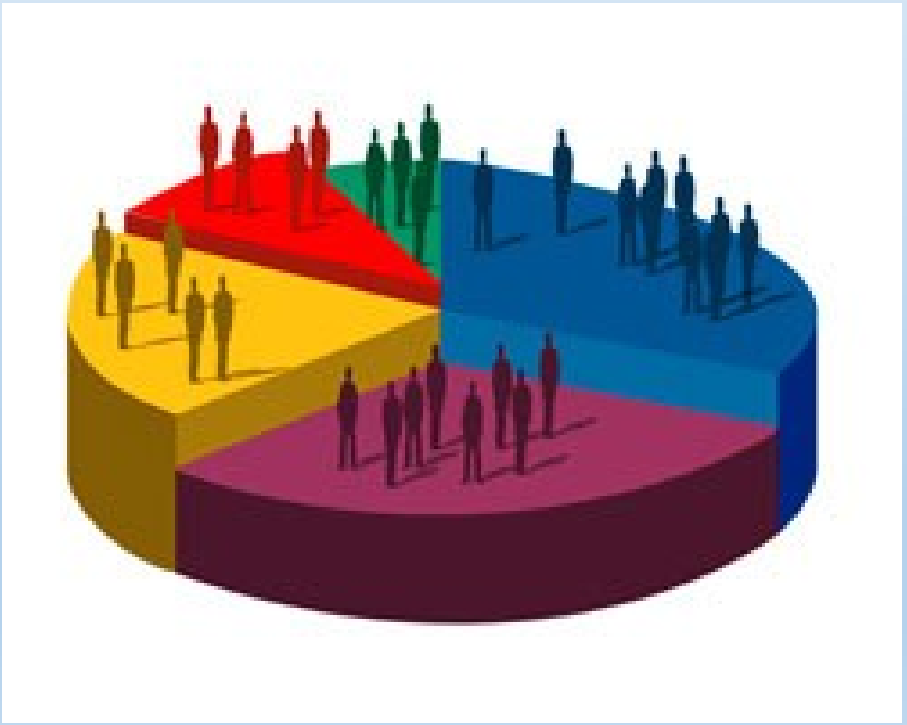
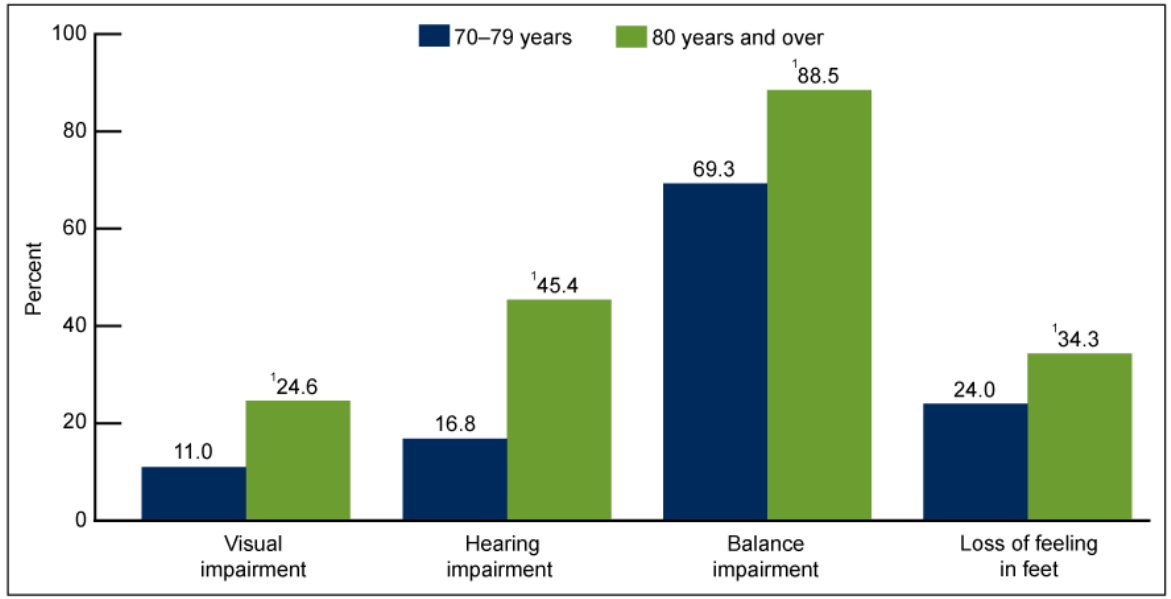


Figure 3. The prevalence of sensory impairments among persons aged 70–79 years compared with persons aged 80 years and over: United States, 1999–2006



[1]Significantly different from the 70–79 age group.

SOURCE: CDC/NCHS, National Health and Nutrition Examination Survey.

Terminologia

L'oggetto dell'osservazione di ogni *fenomeno individuale* che costituisce il *fenomeno collettivo* in studio è detto **'unità statistica'**

Ex: un individuo, una A.S.L., un ospedale...

Ciascuna *unità statistica* presenta delle caratteristiche definite **‘caratteri/variabili’**

Ex: il colore degli occhi in un individuo, il numero di medici che lavorano in una A.S.L., il numero di posti letto in un ospedale....

Ciascun *carattere* è presente in ogni *unità* con una determinata **‘modalità‘**

Ex: colore blu degli occhi, 10 medici in una certa ASL, 300 posti letto in un dato ospedale,...

La classificazione dei caratteri

I caratteri possono essere classificati in:

-Caratteri **qualitativi** (colore degli occhi, tipo di diagnosi...), a loro volta distinti in:

- ***ordinabili*:** è possibile *ordinare* le modalità del carattere in senso crescente o decrescente (es: titolo di studio, livello di gravità della diagnosi...);
- ***sconnessi*:** non c'è alcun ordinamento intrinseco tra le modalità (es: colore degli occhi, sesso...);

-Caratteri **quantitativi** (es: età, peso, numero di medici,...) distinguibili in:

- ***discreti*:** le modalità del carattere sono numeri interi (es: numero di medici...)

- ***continui***: le modalità del carattere sono misurate su una scala continua (es: peso, altezza...).

Alla base di tale classificazione dei caratteri vi è la **'scala di misura'** con cui sono espresse le modalità: se attraverso dei **numeri** o delle '**etichette**'.

SCALA DI MISURA

Nominale (qualitativi sconnessi)

Ordinale (qualitativi ordinabili)

«factor data»

«date/time data»

Discreta (quantitativo)

Continua (quantitativo)

«numeric data»

Approfondimento:

Sintesi delle scale di misura:

Nominale:
Variabili categoriali/qualitative **non ordinate**. Queste possono essere **binarie** (due sole categorie, come il genere: maschio o femmina) o **multinomiali** (più di due categorie, come lo stato civile: sposato, divorziato, mai sposato, vedovo, separato). La cosa fondamentale qui è che **non esiste un ordine logico** per le categorie.

Ordinale:
Categorie **ordinate**. Sempre variabili categoriali/qualitative, ma con un ordine. Gli elementi delle scale «Likert*» con risposte come: "*Mai, a volte, spesso, sempre*" sono ordinali.

Le variabili numeriche/quantitative possono essere ulteriormente suddivise in due tipologie: **discrete** e **continue**. Le variabili discrete, come i **conteggi**, possono assumere solo **numeri interi**: numero di bambini in una famiglia, numero di giorni persi dal lavoro. Le variabili numeriche **continue** possono assumere qualsiasi numero, anche oltre la virgola decimale.

Non è sempre ovvio che questi livelli di misurazione non riguardino solo il significato intrinseco della variabile stessa.

Altrettanto importanti sono il significato della variabile **all'interno del contesto della ricerca** e **come è stata misurata**.

*__Scale likert__: costituite da un certo numero di affermazioni - definiti **item** - che esprimono un atteggiamento positivo e negativo rispetto ad uno specifico tema. La somma di tali giudizi tenderà a delineare l'atteggiamento del soggetto nei confronti del tema. Per ogni item si presenta una scala di accordo/disaccordo, generalmente a 5 o 7 modalità.

Un esempio: l'età

Approfondimento

L'età è una **variabile numerica *continua***. L'età di una persona è *continua* se la si misura con *sufficiente precisione*. È significativo dire che qualcuno (o qualcosa) ha 7.28 anni (?)…

Età come ordinale: **categorie** di età.

Età come discreta: età misurata in **numero di giorni** in cui i semi germinati di una specie iniziano a germogliare. La maggior parte lo farà entro pochi giorni e il valore potrebbe variare da 2 a 9 giorni. In questo contesto, l'età è sicuramente un conteggio discreto: il numero di giorni.

Età come multinomiale: a volte le variabili numeriche sono rese categoriche a causa della mancanza di valori. In uno studio, si misuravano le età di alcuni testimoni in un processo. Tecnicamente le età sono continue, ma in questo studio c'erano solo quattro valori: 49, 59, 70 e 80.

Età come binaria: **sopra/sotto** una certa soglia.

Esempio: studio sull'equilibrio tra lavoro e vita privata di genitori
variabile di interesse: numero di ore lavorate a settimana.
Predittore chiave: età del figlio più piccolo.

[C'è una differenza tra un bambino di 5 anni, che può essere idoneo solo per la scuola materna part-time e un bambino di 6 anni, che è abbastanza grande per andare a scuola a tempo pieno].

Esempio di **matrice dei dati**:

Abbiamo chiesto ad un campione di studenti se avessero visto l'ultimo Festival di Sanremo e se lo avessero gradito. Inoltre, abbiamo chiesto il **genere** (maschio/femmina) e il **numero di ore giornaliere** dedicate al sonno e ad attività di studio.

variabile ↓

Sappiamo definire la **scala di misura** di queste variabili?

Stu.	sesso	Sanremo	⋯	sonno	studio
1	maschio	Non l'ho visto	⋯	8	2
2	femmina	L'ho visto e mi è piaciuto	⋯	6	30
3	maschio	Non l'ho visto	⋯	9	5
4	femmina	Non l'ho visto	⋯	8	25
⋮	⋮	⋮	⋮	⋮	⋮
52	femmina	Non l'ho visto	⋯	8	20

← *unità statistica*

Classificazione dei caratteri e scala di misura

CARATTERE		SCALA
qualitativo	Sconnesso	Nominale
	Ordinabile	Ordinale
quantitativo		Ad intervalli (scala numerica discreta o continua)

Operazioni che è possibile fare sui caratteri in base alla loro scala di misura:

Operazioni sulle modalità del carattere	**Carattere**		
	qualitativi		**Quantitativi**
	sconnessi	*ordinabili*	*(discreti/continui)*
= ; ≠	si	si	si
> ; <	no	si	si
+ ; -	no	no	si

Distribuzioni

Quando si rilevano le **modalità** con cui uno (o piu') **caratteri** si presentano in ciascuna **unità** di una popolazione (o di un campione) si ottiene una **'distribuzione'** della popolazione/campione secondo il carattere considerato.

Ex: distribuzione 'unitaria' del peso in una classe:

Studente	Peso (kg)
Marco	62
Chiara	59
Luca	56
...	...

distribuzione «**unitaria**» = ad ogni unità del gruppo è associato il suo peso, cioè la corrispondente modalità del carattere considerato.

Possiamo suddividere in **'classi'** la popolazione secondo il carattere considerato, allora le modalità del carattere vengono raggruppate in classi ed otteniamo una distribuzione di **'frequenze' ->** *frequenza della classe* = numero di unità statistiche che appartengono alla classe.

Distribuzioni di frequenze

Tipo di scuola

		Frequenza	Percentuale
Validi	Altro	11	34,4
	Istituto Professionale	4	12,5
	Liceo Classico	3	9,4
	Liceo Scientifico	14	43,8
	Totale	32	100,0

Il numero di unità che appartengono ad una classe= **frequenza assoluta** della classe.

E' possibile calcolare anche le distribuzioni di frequenze relative e di frequenze percentuali

Alcuni simboli:

Abbiamo *k* classi di un dato carattere (nell'ex: 6 classi) cioè *k*=6; le possiamo indicare tramite un indice *i* che varia da 1 a 6: *i*=1,2,3,4,5,6. Abbiamo rilevato le modalità del carattere su *N* unità (nell'ex: *N*=30).

- Frequenza **assoluta** della classe *i* -> n_i
- Frequenza **relativa** della classe *i* -> $f_i=n_i/N$
- Frequenza **percentuale** della classe *i* -> $p_i=f_i*100=(n_i/N)*100$

Esiste poi un simbolo matematico indica l'operazione di **somma** su un gruppo di oggetti ed è chiamato '**sommatoria**' :

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 30$$

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6$$

(per *k*=6)

$$\sum_{i=1}^{k} n_i = N \; ; \quad \sum_{i=1}^{k} f_i = 1 \; ; \quad \sum_{i=1}^{k} p_i = 100$$

Distribuzioni doppie (dati categorici/in classi)

Se si rilevano **due** caratteri in una popolazione si ottiene una **distribuzione *doppia***; è uso comune rappresentare le distribuzioni doppie delle modalità rilevate per due caratteri (**categorici o espressi in classi**) tramite la **'tabella di contingenza**' o **'tabella a doppia entrata**‘:

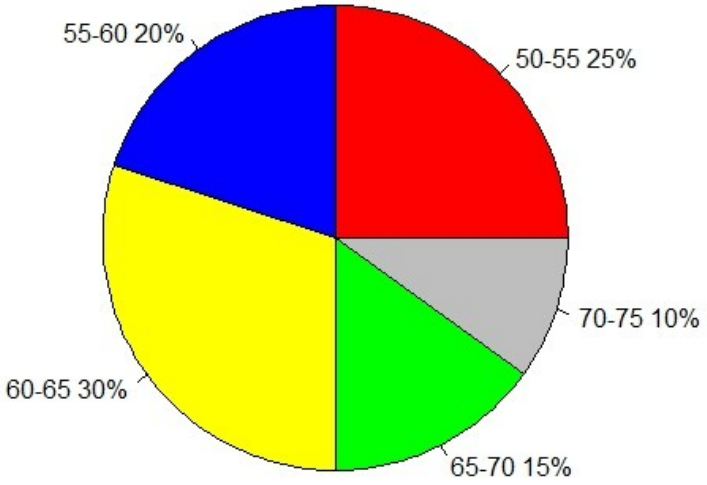
CLASSI DI PESO	M	F	TOT
45-50	0	2	**2**
50-55	5	2	**7**
55-60	4	3	**7**
60-65	6	2	**8**
65-70	3	1	**4**
70-75	2	0	**2**
TOT	**20**	**10**	**30**

Distribuzioni doppie (dati categorici/in classi)

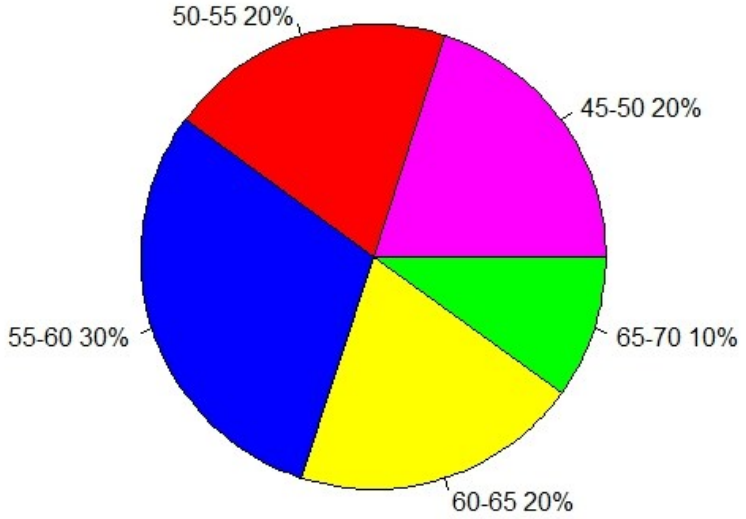
Per confrontare la distribuzione delle classi di peso rispetto al sesso occorre calcolare le **frequenze relative** all'interno della tabella di contingenza:

CLASSI DI PESO	M	F
45-50	0	20
50-55	25	20
55-60	20	30
60-65	30	20
65-70	15	10
70-75	10	0
TOT	100	100

MASCHI



FEMMINE



La tabella di contingenza permette inoltre di studiare **l'associazione** fra i caratteri analizzati
→ **test di ipotesi** che vedremo in seguito → **Statistica Inferenziale**

Tabella di Contingenza (a doppia entrata):

n_{aibj} = numero di soggetti che presentano *contemporaneamente* la modalità *i* del carattere A e la modalità *j* del carattere B.

		B				Totale
		B_1	B_2	...	B_m	
A	A_1	na_1b_1	na_1b_2	...	na_1b_m	NA_1
	A_2	na_2b_1	na_2b_2	...	na_2b_m	NA_2
	...	...	...	na_ib_j	...	...
	...	...	...	...	...	...
	A_k	na_kb_1	na_kb_2	...	na_kb_m	NA_k
Totale		NB_1	NB_2	...	NB_m	N

Marginali di Riga

Marginali di colonna

Modalità	Frequenza assoluta
160	2
164	2
165	1
166	1
168	1
170	5
172	1
173	4
174	1
175	4
176	4
177	1
178	2
180	6
181	2
182	2
183	3
184	2
185	3
186	1
187	2
188	1
190	2

Abbiamo raccolto dei dati sulla distribuzione delle altezze in un campione di 53 studenti.
Vogliamo suddividere questa distribuzione in **classi di frequenza**.
Con che criterio scegliamo le classi?

Modalità	Frequenza assoluta
[160,165]	5
(165,170]	7
(170,175]	10
(175,180]	13
(180,185]	12
(185,190]	6

Possiamo creare ad esempio
6 classi di frequenza:
[: la classe comprende il valore;
(: la classe NON comprende il valore;

→ **Classi [circa] di uguale ampiezza**

Modalità	Frequenza assoluta
[160,185]	47
(185,190]	6

Ma possiamo creare ad esempio
solo 2 classi di frequenza:
[: la classe comprende il valore;
(: la classe NON comprende il valore;

La scelta delle classi è arbitraria, ma va fatta in maniera ragionevole….

Può capitare, o per scelta (si vuole fornire informazioni più dettagliate su una parte della distribuzione) o per necessità (i dati ci sono stati forniti già raggruppati in classi e non disponiamo della distribuzione unitaria) di costruire delle classi utilizzando ***intervalli di lunghezza differente***.

In questo caso è conveniente definire il concetto di **densità** di frequenza della classe:

Densità=frequenza assoluta classe / ampiezza della classe

La densità ci dice il *numero atteso di unità statistiche* per ogni unità di misura della variabile.

Esempio: Numero di amici su Facebook, indagine fatta su un campione di 52 studenti.

Modalità	Freq. ass.
0	1
11	1
79	1
100	1
112	1
119	1
130	2
140	1
150	1
162	1
169	1
176	1
200	2
231	1
254	1
257	1
277	1
300	1
349	1
350	1
356	2
370	1
400	1
438	1
439	1
450	1
463	1
469	1
470	1
500	2
520	1
543	2

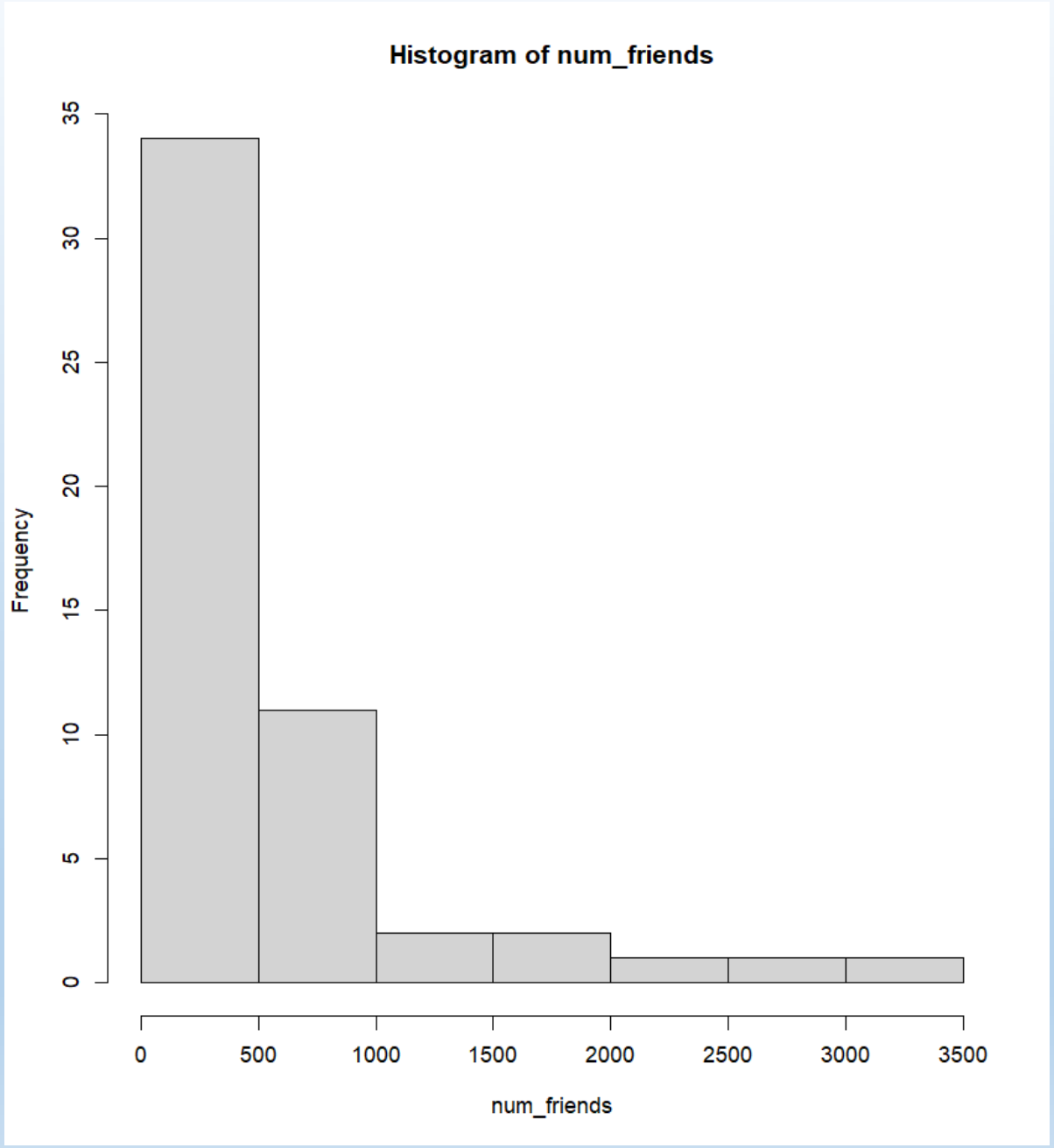
….

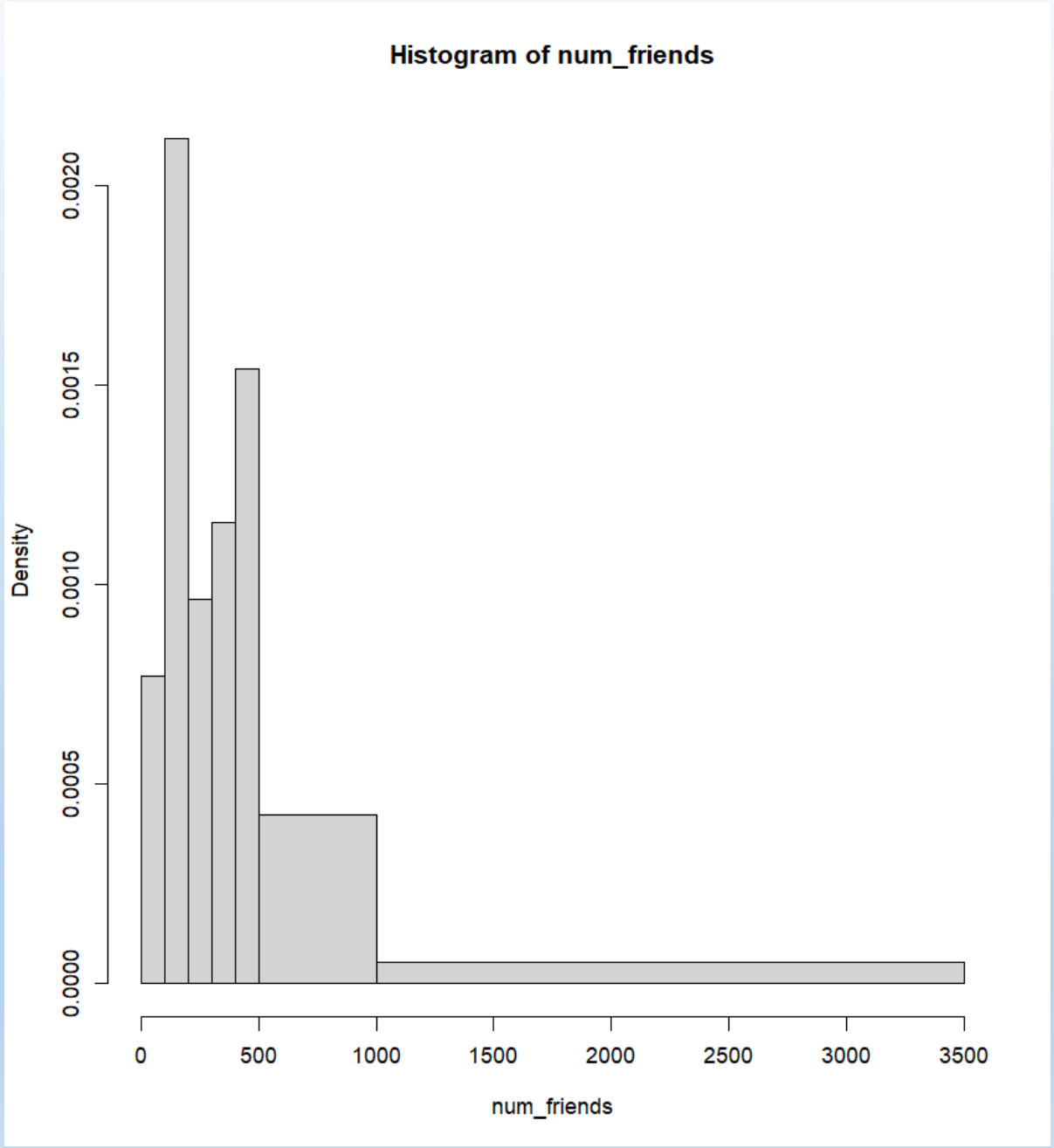
Modalità	Frequenza assoluta	Ampiezza Classe	Densità	"Altezza" della densità (grafico) Freq rel/ampiezza
[0,100]	4	100	0,04	**(4/52)/100=0.0008**
(100,200]	11	100	0,11	**0.0021**
(200,300]	5	100	0,05	**0.0010**
(300,400]	6	100	0,06	**0.0012**
(400,500]	8	100	0,08	**0.0015**
(500,1000]	11	500	0,022	**0.0004**
(1000, 3500]	7	2500	0,0028	**0.0001**

La densità ci dice il *numero atteso di unità statistiche* per ogni unità di misura della variabile.

Nella classe [0,100] **ci aspettiamo di osservare 4 persone** in un **intervallo di 100 unità**.

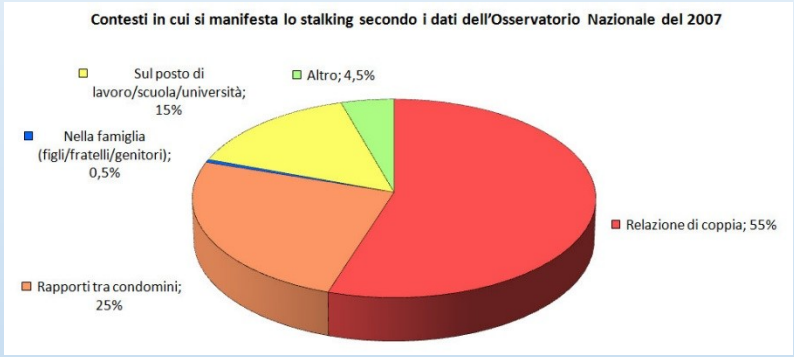
Nella classe (500,1000] ci aspettiamo di vedere 2.2 persone ogni 100 unità (cioè 2.2 persone tra 500 e 600, altrettante tra 600 e 700, …etc).


Histogram of num_friends
Frequency
0
5
10
15
20
25
30
35
0
500
1000
1500
2000
2500
3000
3500
num_friends


Histogram of num_friends
Density
0.0000
0.0005
0.0010
0.0015
0.0020
0
500
1000
1500
2000
2500
3000
3500
num_friends

Rappresentazioni grafiche dei dati

Una parte molto importante della statistica descrittiva è la **rappresentazione grafica** dei dati

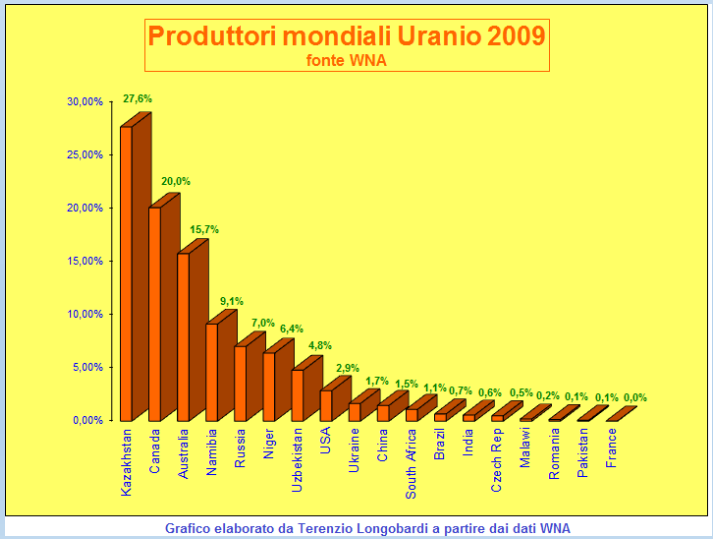


1. Scoprire la **struttura** dei dati;

2. Identificare **i valori più rilevanti** assunti dal fenomeno;

3. Identificare eventuali **valori anomali** (*outliers*)

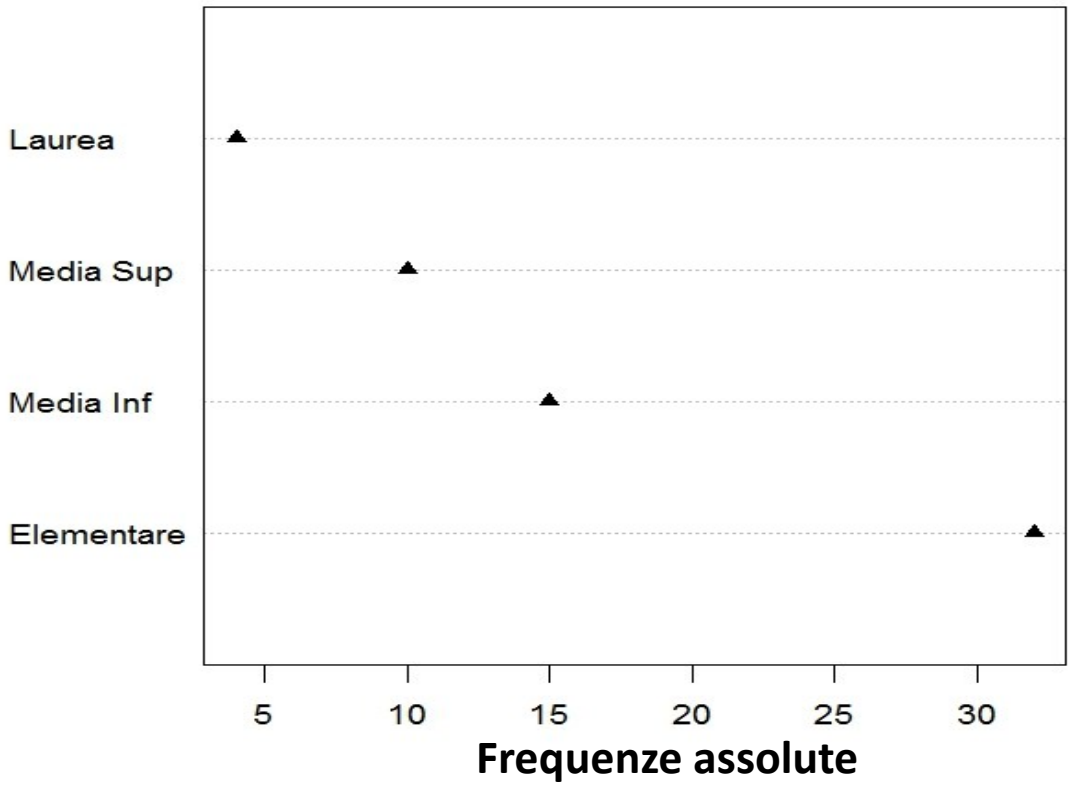
4. **Comunicare** i risultati dello studio in modo efficace

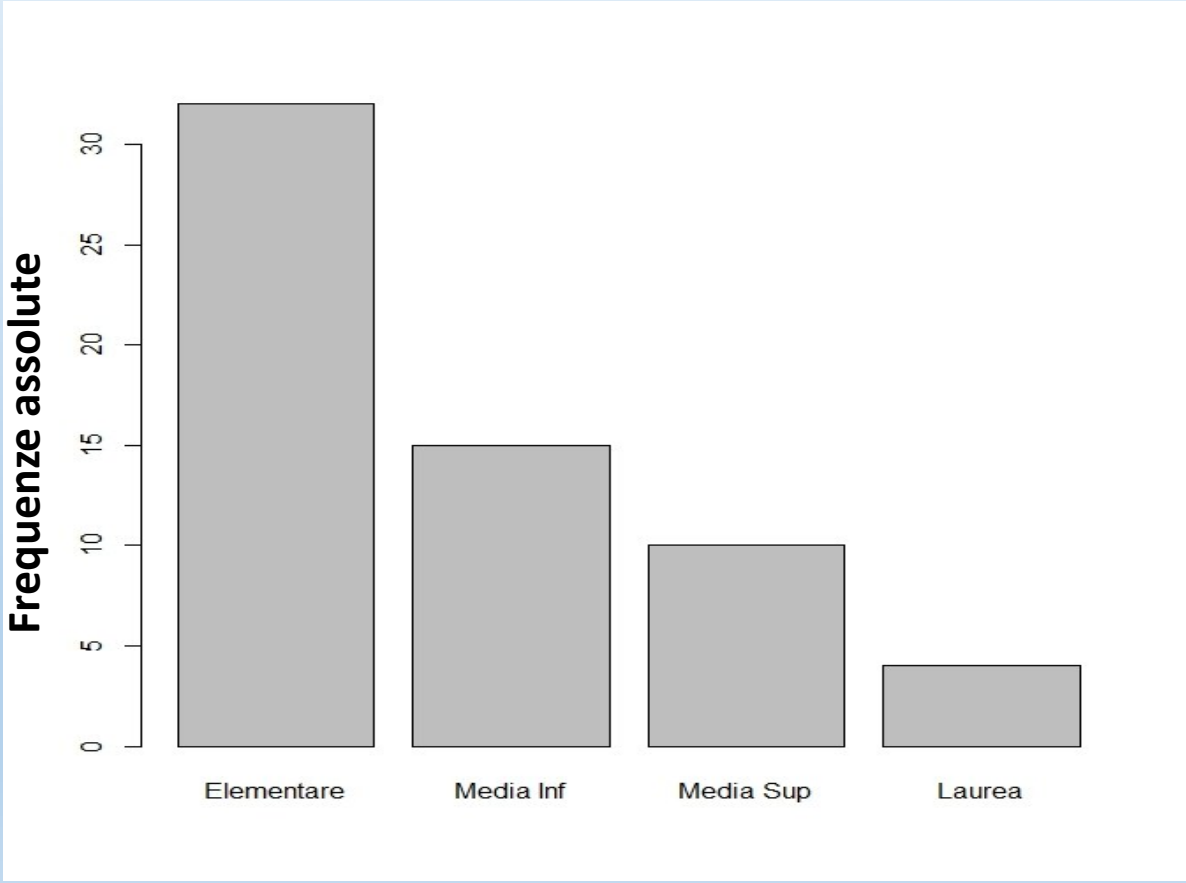


Caratteri su scala nominale/ordinale (I):

Per rappresentare graficamente una distribuzione secondo un carattere di tipo nominale o ordinale possiamo utilizzare il grafico «dotchart» oppure il «barplot»:

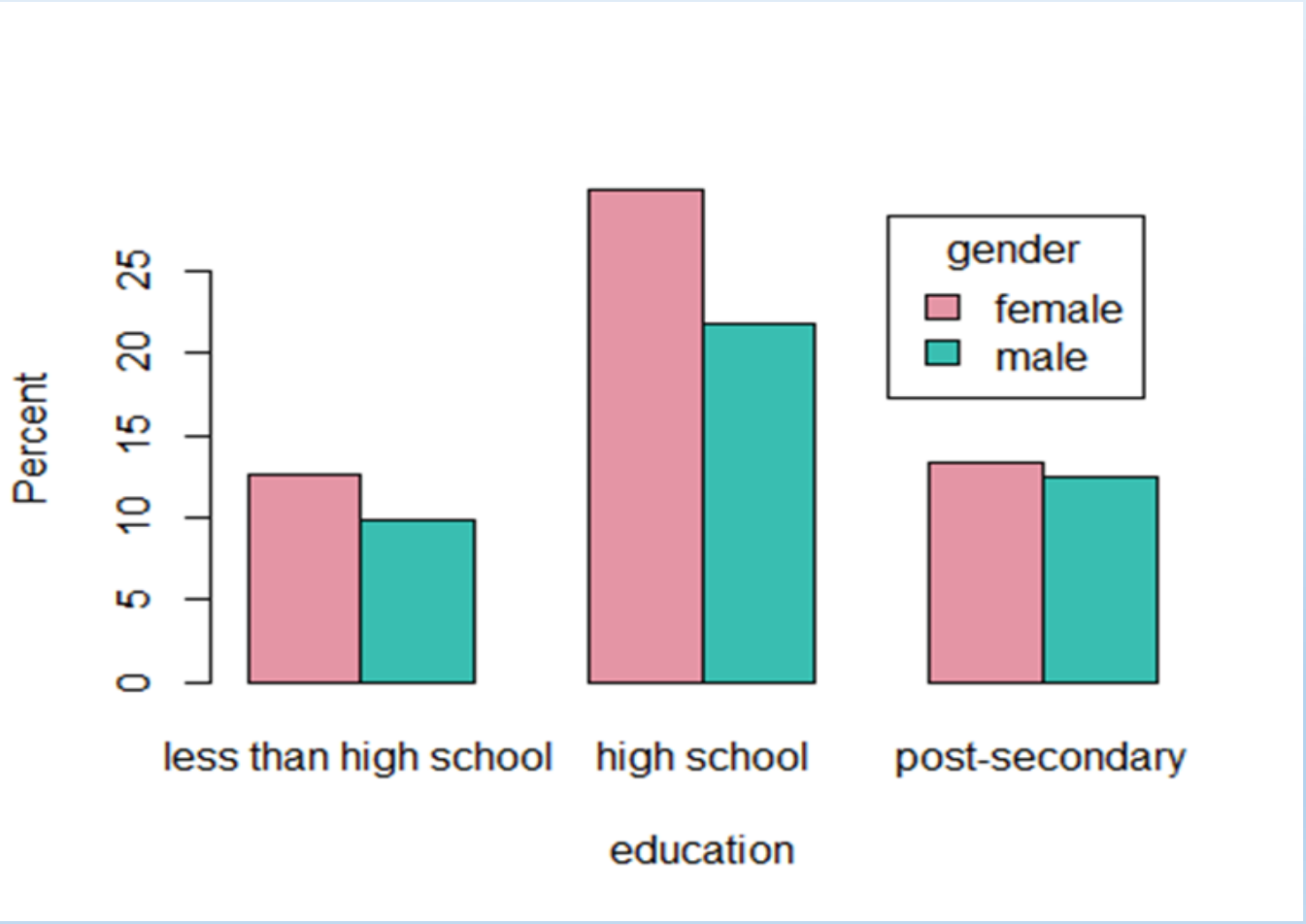
Elementare	Media Inf	Media Sup	Laurea
32	15	10	4





Caratteri su scala nominale/ordinale (II):

Per rappresentare graficamente una **distribuzione doppia** di due caratteri nominali/ordinali possiamo utilizzare il grafico «barplot» [meglio utilizzare le frequenze percentuali]:

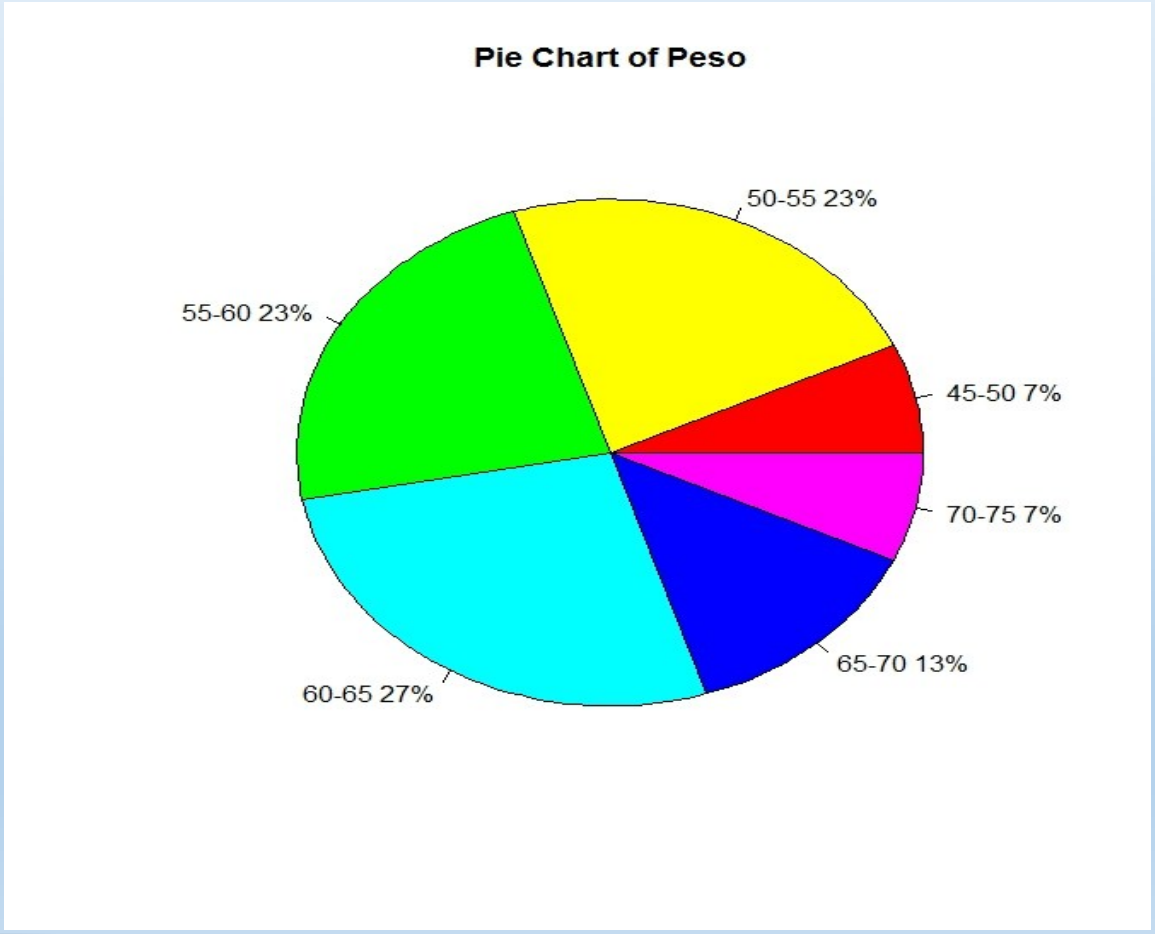


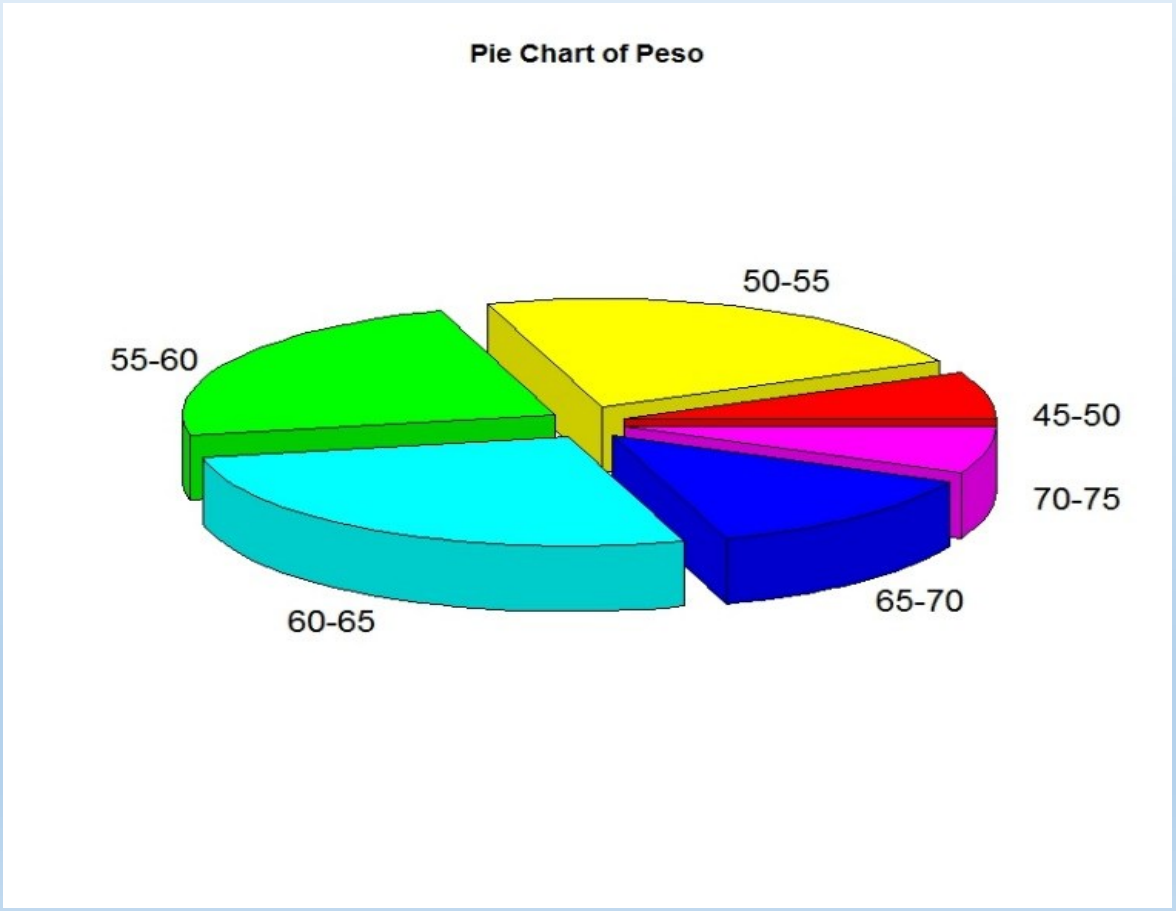

```
Frequency table:
                          gender
education                  female male
  high school                9987 7264
  less than high school      4228 3280
  post-secondary             4436 4159

Total percentages:
                          female male Total
high school                 29.9 21.8  51.7
less than high school       12.7  9.8  22.5
post-secondary              13.3 12.5  25.8
Total                       55.9 44.1 100.0
```

Rappresentazione grafica di distribuzioni in %

Per rappresentare una distribuzione in percentuale secondo un carattere di tipo categorico/in classi possiamo anche utilizzare il grafico a torta:



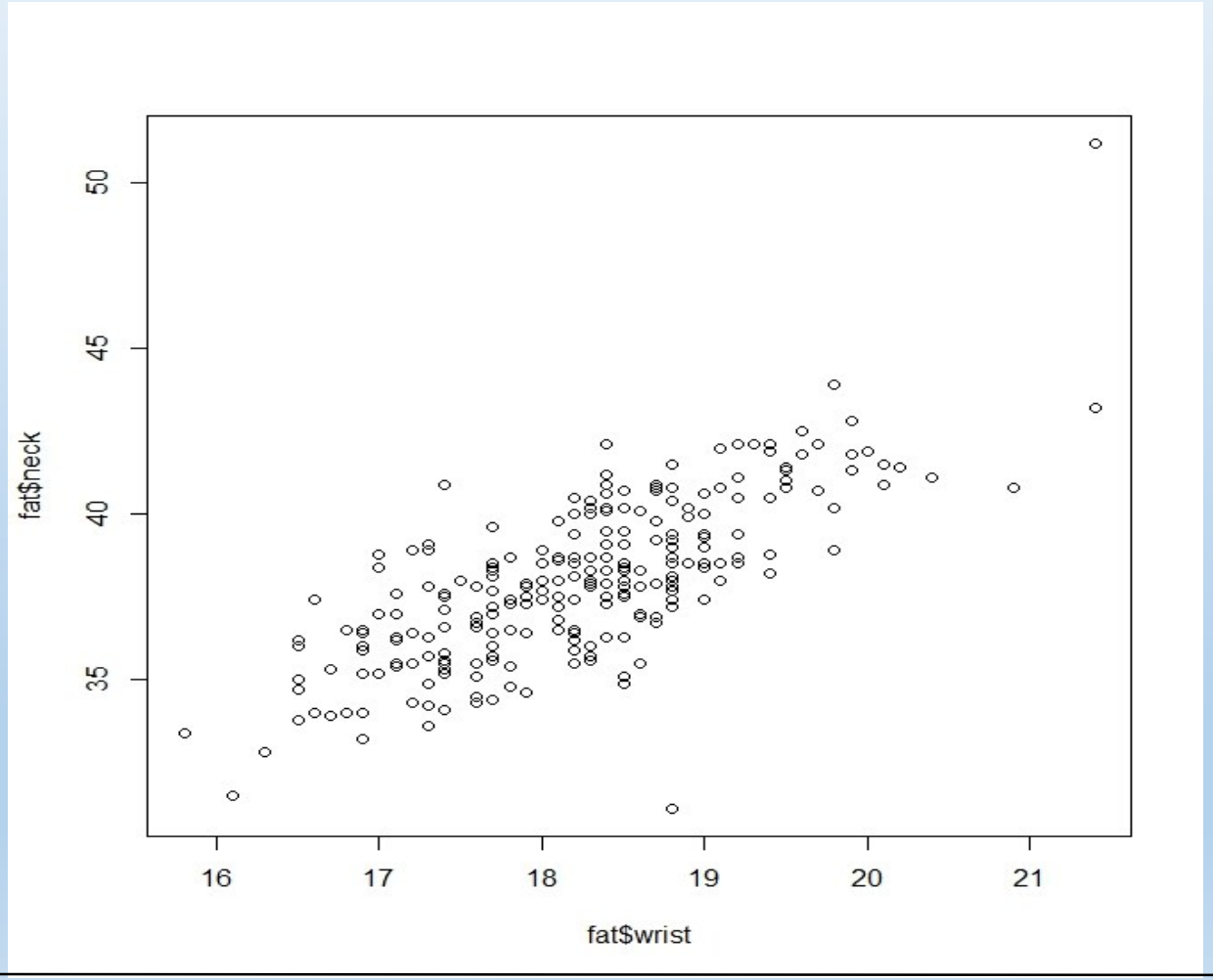


Distribuzioni doppie (dati su scala numerica)

Se si rilevano **due** caratteri in una popolazione su scala numerica e si è interessati a studiare l'associazione tra i due caratteri, la rappresentazione grafica usuale è lo **scatter plot (diagramma cartesiano)**:

Distribuzione su un campione di 252 maschi delle dimensioni del polso e del collo: si ipotizza che l'ampiezza del collo sia circa due volte quella del polso.

Vedremo in seguito come sintetizzare da un punto di vista statistico queste relazioni (**correlazione**/**regressione lineare…**)



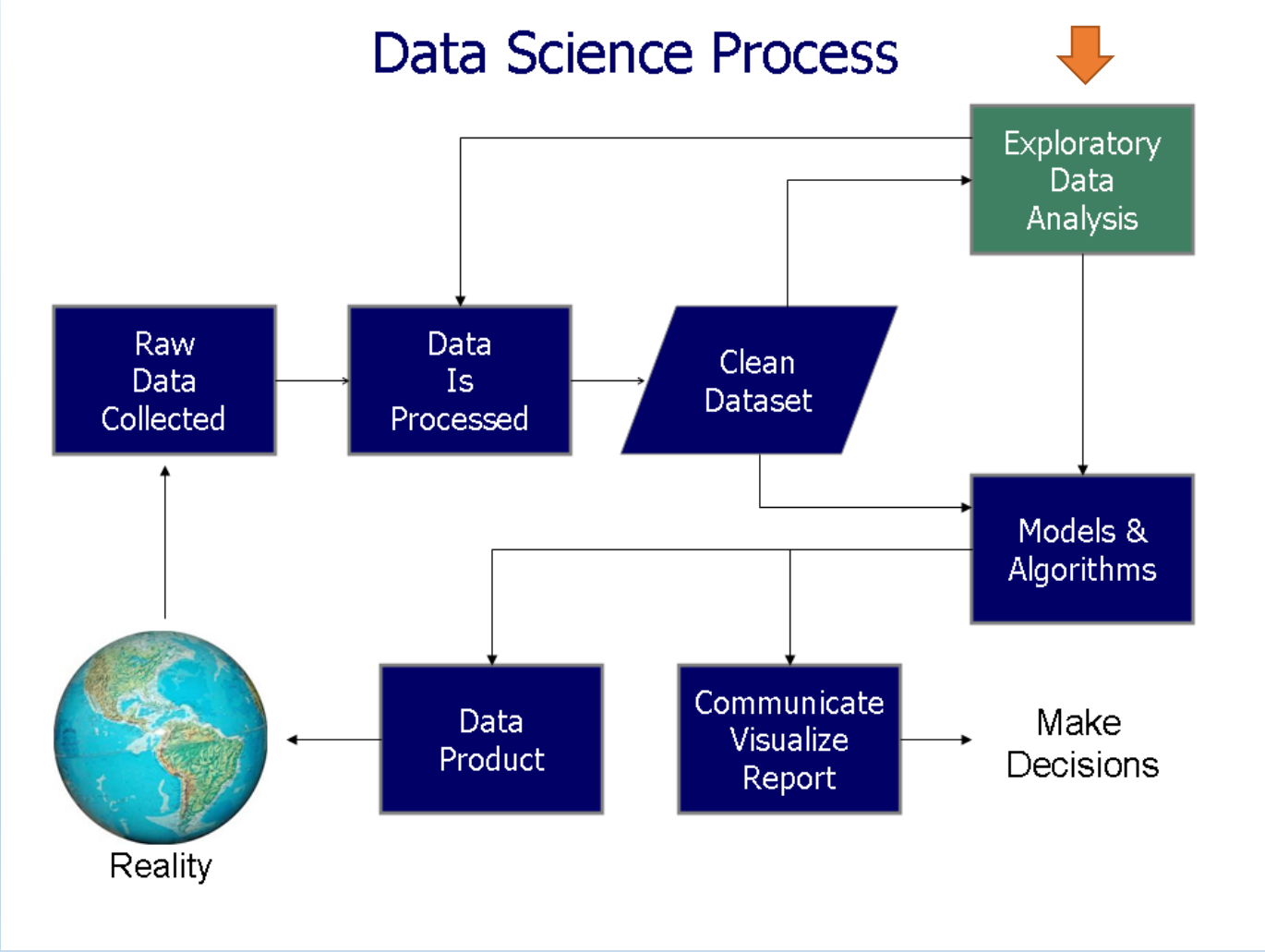
- Indici di Posizione

- Indici di Dispersione

Statistics is the grammar of science.

Karl Pearson (1857-1936).

WE ARE HERE



Gli indici di posizione delle distribuzioni

Per rappresentare in modo obiettivo una massa di informazioni un indice sintetico deve essere facilmente comprensibile, relativamente semplice da calcolare e soprattutto confrontabile con indici ricavati in tempi e luoghi diversi, sullo stesso tipo di dati.

Il dato di sintesi deve essere compreso tra il valore piu' piccolo e quello piu' grande tra quelli osservati (se è possibile ordinarli); deve identificarsi, in qualche modo, con i valori piu' frequenti, i quali corrispondono spesso a quelli localizzati *al centro* delle misure *ordinate*.

Si definiscono quindi gli **'indici di tendenza centrale'** o di **'indici di posizione'**. Un corretto approccio al problema richiede un'analisi preventiva della scala di misura con cui sono espresse le modalità del carattere al fine di scegliere **il tipo di indice** di posizione opportuno da utilizzare.

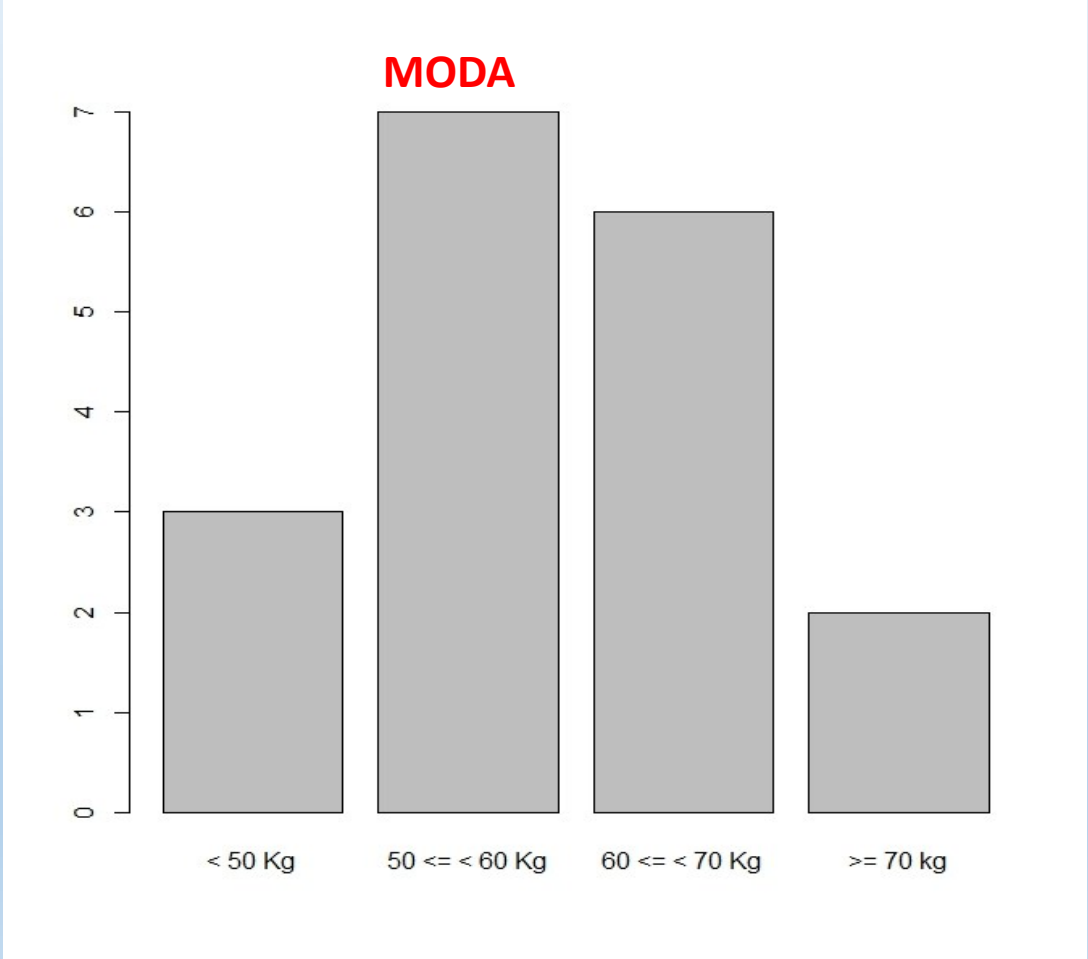
Gli indici di posizione delle distribuzioni (I): MODA

MODA:

Dato un qualsiasi tipo di carattere (qualitativo o quantitativo) la **MODA** della popolazione distribuita secondo quel carattere è la **modalità prevalente** del carattere, ossia quella a cui è associata la massima frequenza.

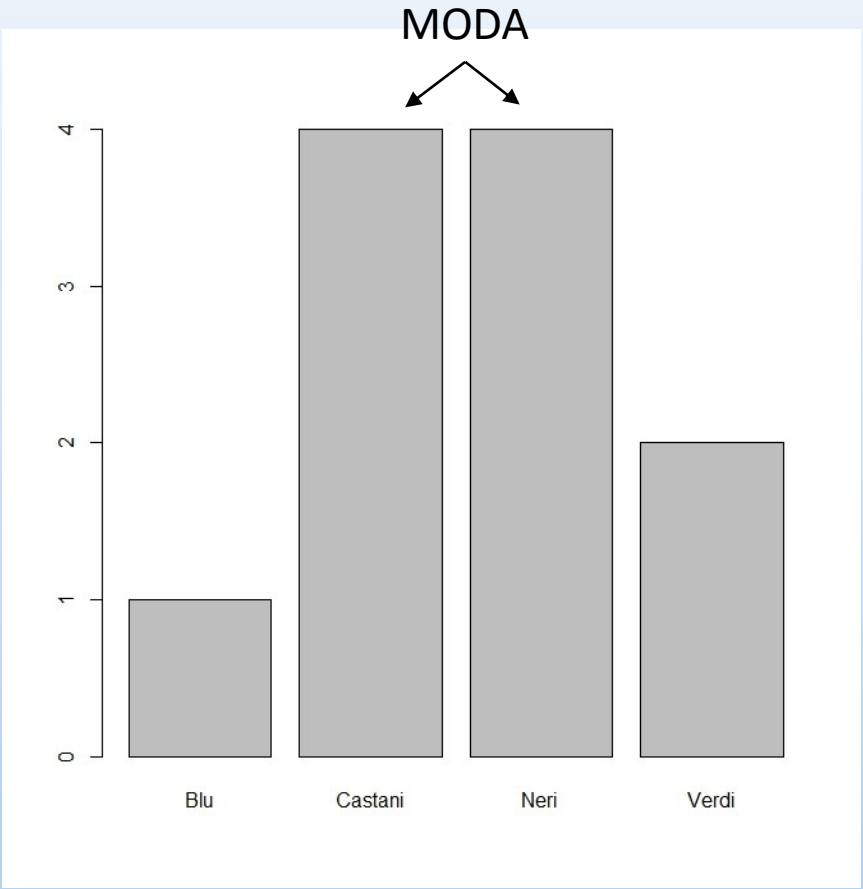
Classe di peso (kg)	**Frequenza di studenti nella classe di peso**
< 50	3
50 ≤ < 60 (Moda o Classe Modale)	**7**
60 ≤ <70	6
≥ 70	2
Tot	18

Non è detto che vi sia un'unica moda in una distribuzione!



Gli indici di posizione delle distribuzioni (I): MODA

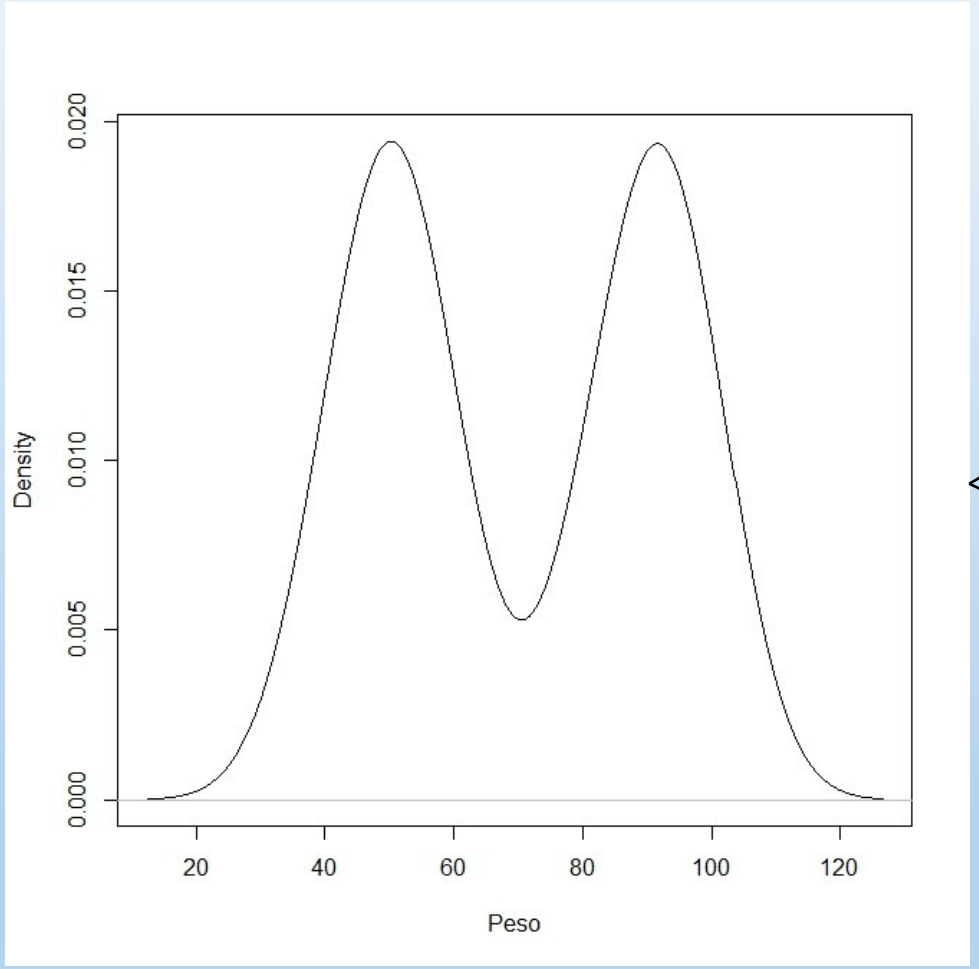
Colore degli occhi	Frequenza di studenti con quel colore di occhi
Blu	1
Castano (Moda o Classe Modale)	**4**
Nero (Moda o Classe Modale)	**4**
Verde	2
totale	11

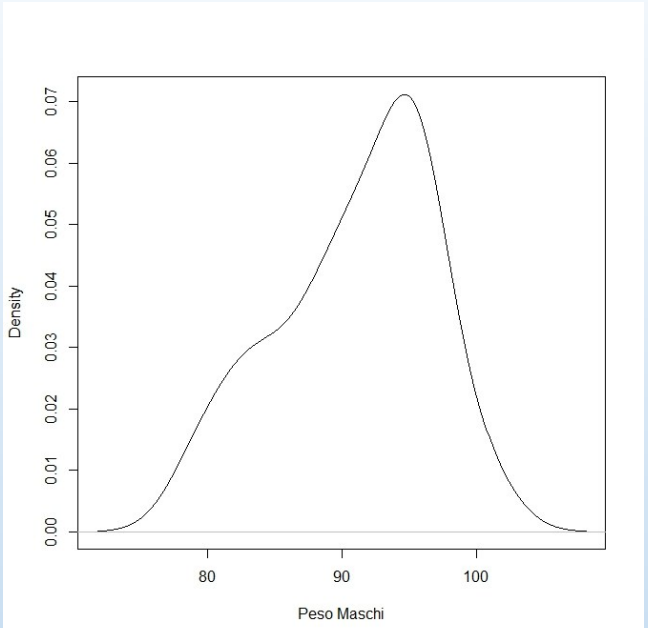


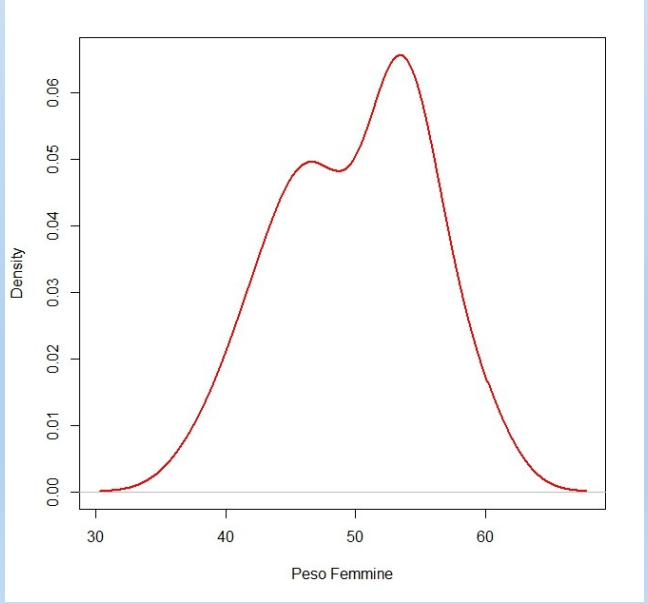
…in questo caso la distribuzione viene definita *bi-modale*, cioè ha due mode.

Per quanto riguarda i **caratteri qualitativi sconnessi** la moda è l'**unico** indice
di posizione che si puo' calcolare.

Gli indici di posizione delle distribuzioni (I): MODA



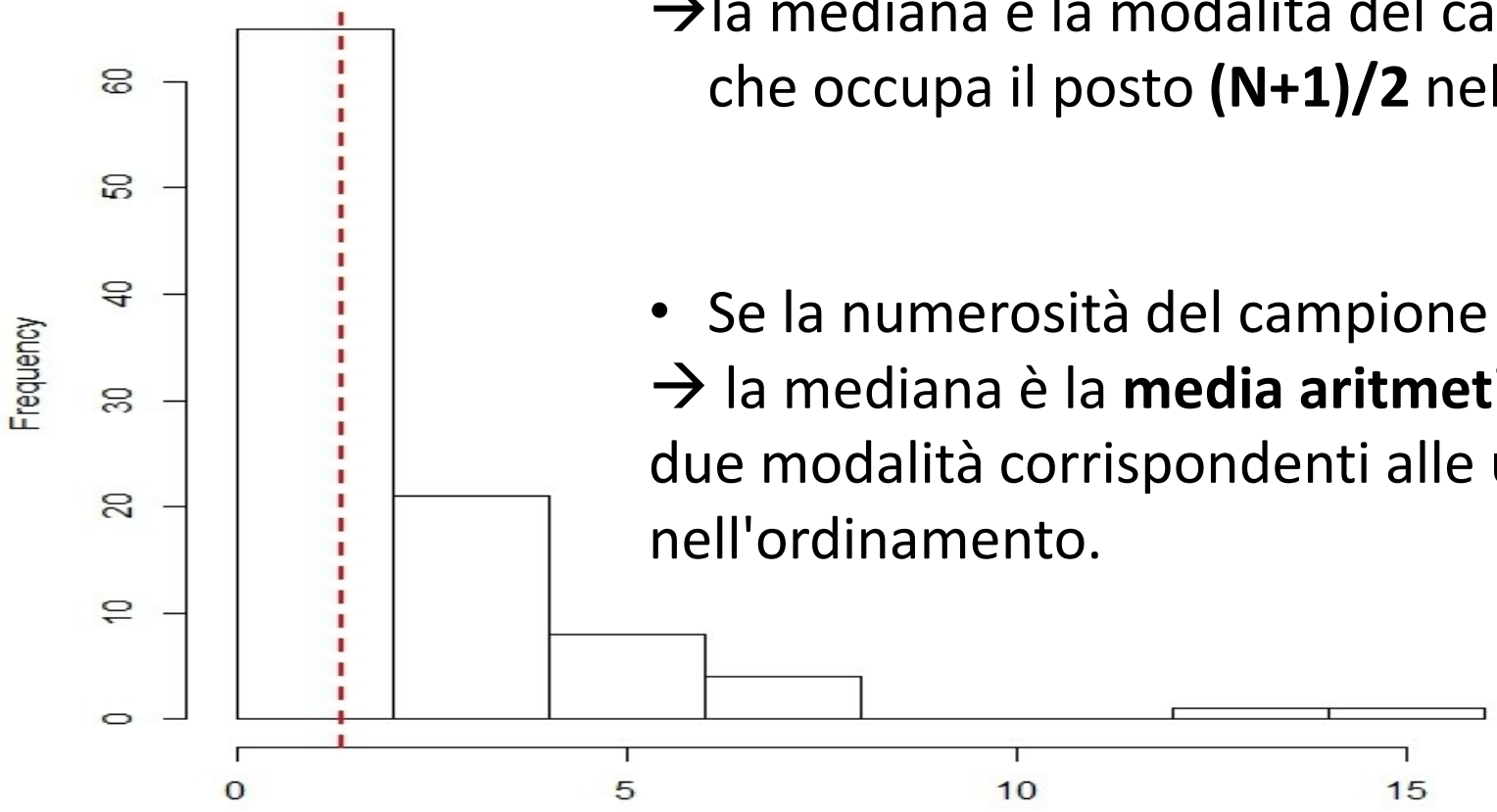




La moda ci aiuta a capire se la distribuzione è **omogenea** oppure no!

Gli indici di posizione delle distribuzioni (II): MEDIANA

Dato un **carattere ordinabile** la **MEDIANA** della distribuzione è la modalità del carattere che bi-partisce (=divide in *due parti uguali*) la distribuzione.



- Se la numerosità del campione **N** è dispari,

→la mediana è la modalità del carattere associata all'unità che occupa il posto **(N+1)/2** nell'ordinamento;

- Se la numerosità del campione **N** è pari,

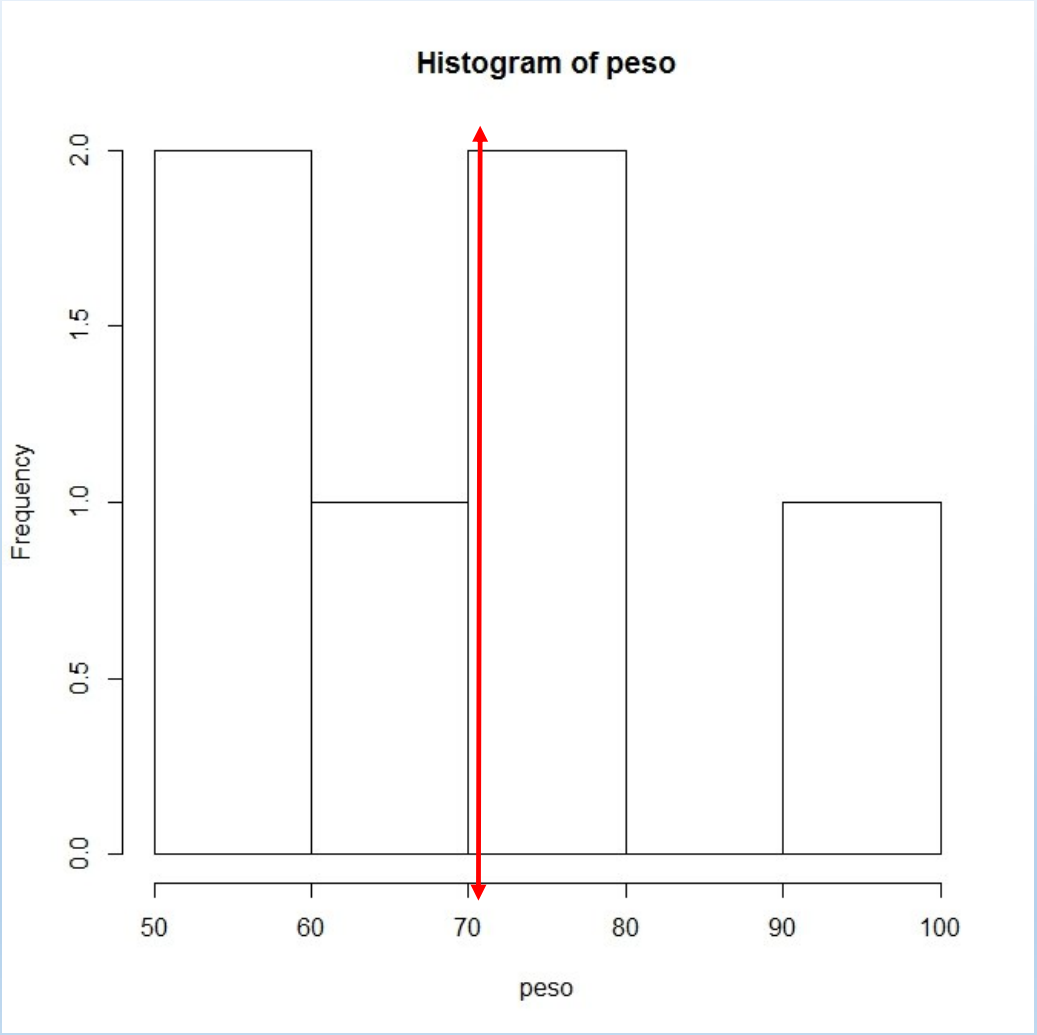
→ la mediana è la **media aritmetica** dei valori assunti dalle due modalità corrispondenti alle unità centrali nell'ordinamento.

Gli indici di posizione delle distribuzioni (II): MEDIANA

SOGGETTO	PESO (kg)	Rango
SOGGETTO 1	55	2
SOGGETTO 2	78	5
SOGGETTO 3	52	1
SOGGETTO 4	**67**	**3**
SOGGETTO 5	91	6
SOGGETTO 6	**76**	**4**

→ la mediana è: (67+76)/2=71,5 Kg

Il 50% del campione esaminato ha un peso corporeo inferiore o uguale a 71,5 Kg (*sintesi* dei dati).

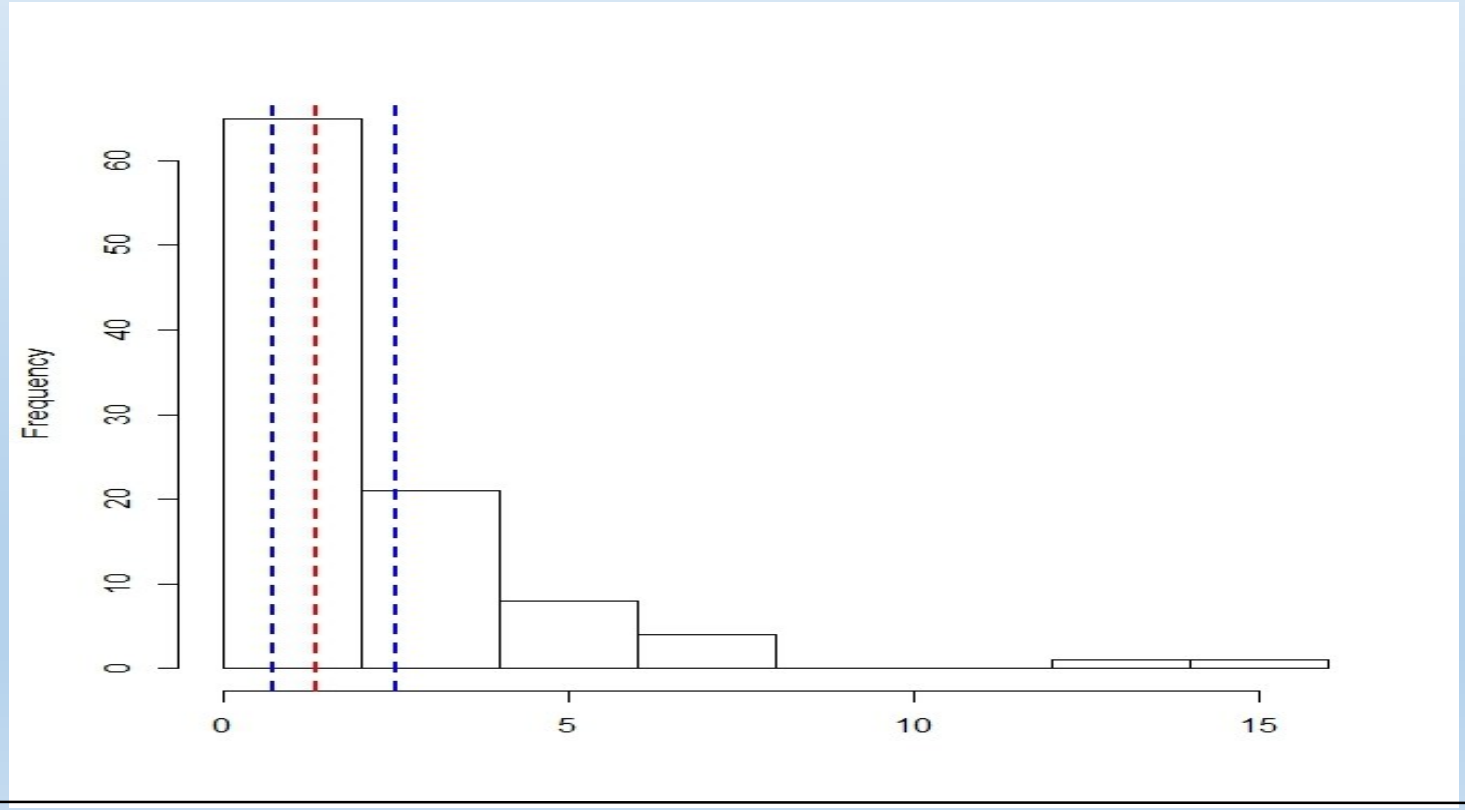


Gli indici di posizione delle distribuzioni (III): Quantili/Percentili

I '**quantili**' o '**percentili**' identificano alcuni indici di posizione che altro non sono che un'estensione del concetto di mediana: suddividono in parti uguali una serie ordinata di dati.

I '**quartili**' sono gli indici di posizione che dividono una serie ordinata di dati in **4** parti uguali:

Q1= primo quartile:
25% della distribuzione
Q3= terzo quartile:
75% della distribuzione



Nel caso di caratteri qualitativi ordinabili o quantitativi è possibile definire accanto alle frequenze assolute e relative (e percentuali), le **frequenze cumulate assolute e relative** (e percentuali).

Considerata una distribuzione di frequenze, siano $x_1, x_2, \dots, x_K$ le **modalità** assunte (ordinate in ordine crescente) da un carattere qualitativo ordinabile o quantitativo sulle N unità di un collettivo:

X	n_i	f_i	N_i	F_i
x_1	n_1	f_1	$N_1 = n_1$	$F_1 = f_1$
x_2	n_2	f_2	$N_2 = n_1+n_2$	$F_2 = f_1+f_2$
x_i	n_i	f_i	$N_i = n_1+n_2+\dots+n_i$	$F_i = f_1+f_2+\dots+f_i$
x_K	n_K	f_K	$N_K = n_1+n_2+n_i+\dots n_k = N$	$F_K = f_1+f_2+f_i+\dots f_k = 1$
Totale	**N**	**1**		

La frequenza assoluta cumulata N_i misura quante unità del collettivo osservato possiedono o la modalità x_1 o la modalità x_2 … o «fino alla» modalità x_i.

Per i=1 abbiamo che N_1 è esattamente uguale alla frequenza assoluta della prima modalità.

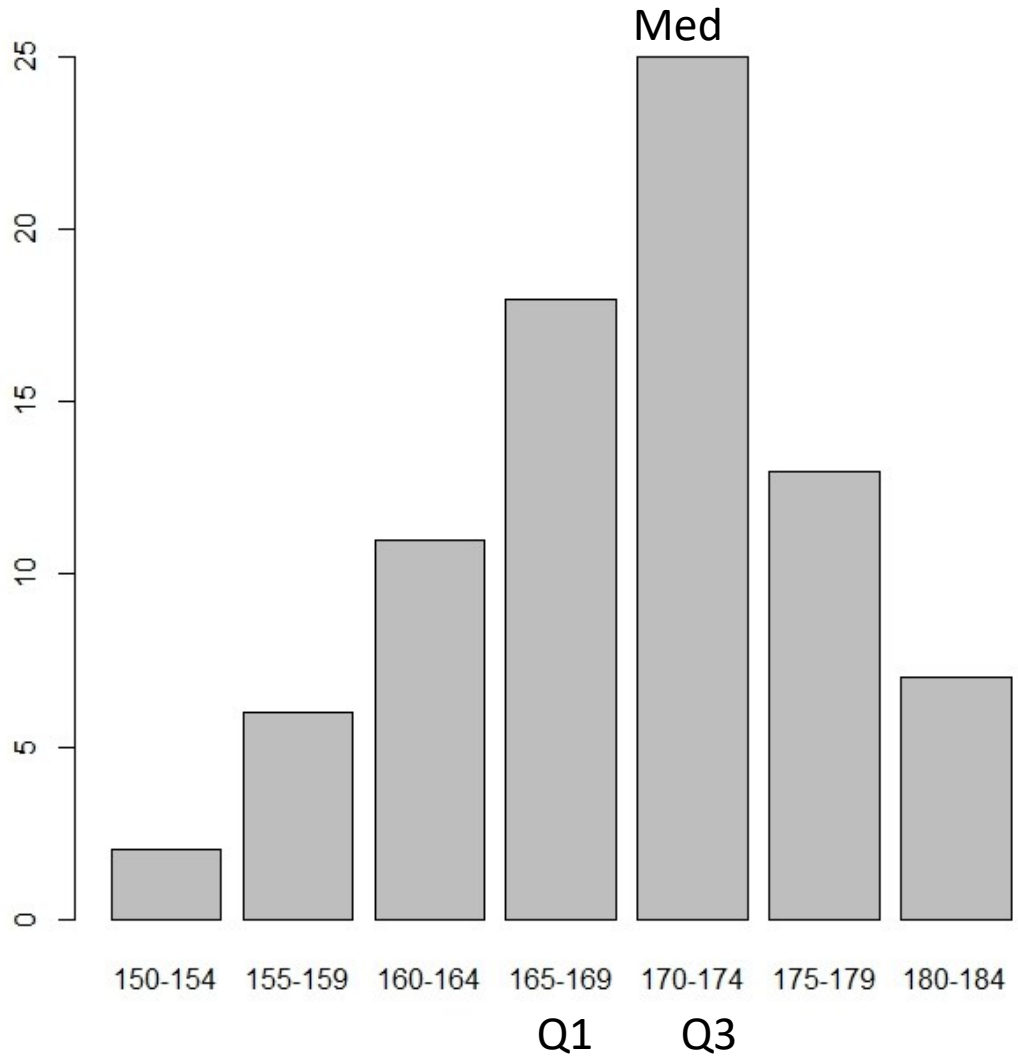
Per i=K abbiamo che N_K è uguale a tutta la numerosità del collettivo (N).

FREQUENZE CUMULATE ASSOLUTE

FREQUENZE CUMULATE RELATIVE: N_i/N

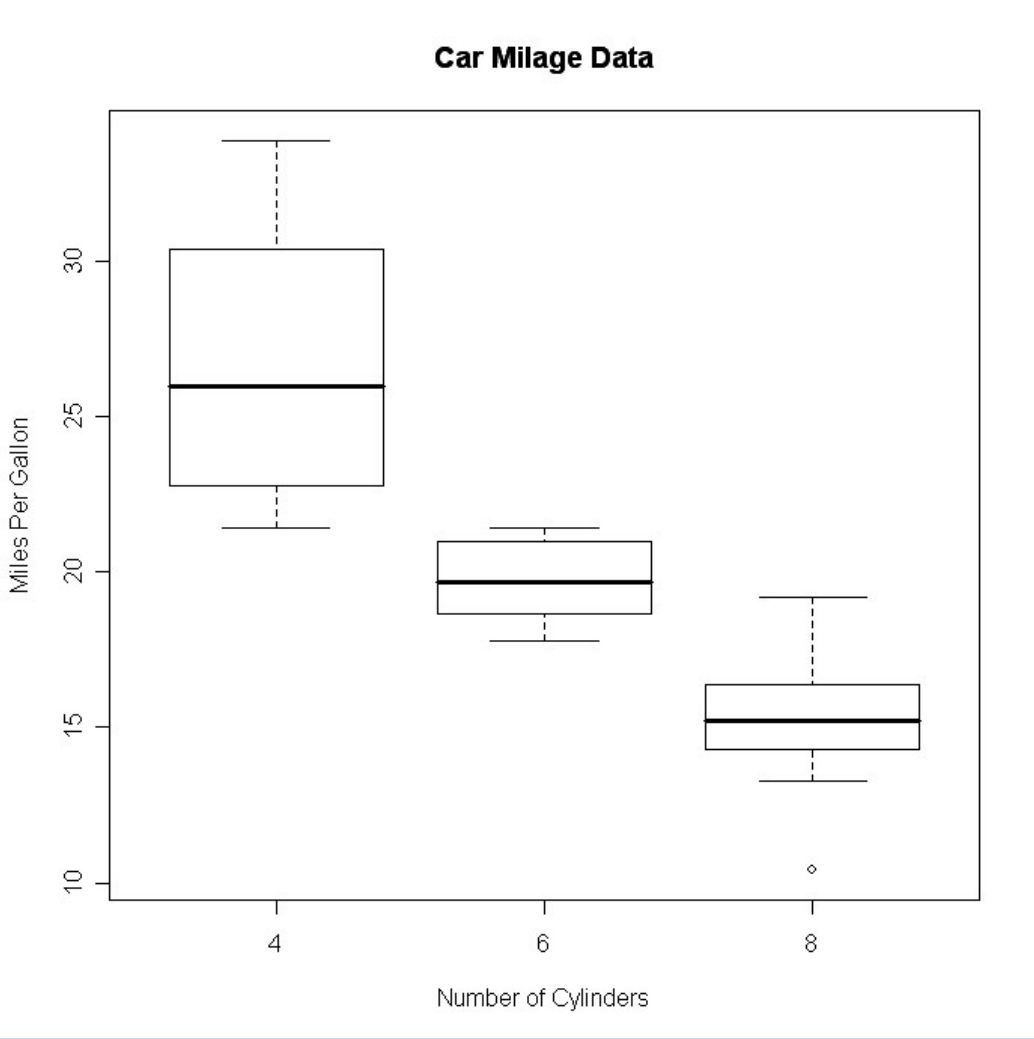
Gli indici di posizione delle distribuzioni (III): Quantili/Percentili

Classi di altezza (cm)	Frequenze assolute	Frequenze cumulate assolute	Frequenze cumulate percentuali (%)
150 – 154	2	2	2
155 – 159	6	8	10
160 – 164	11	19	23
165 – 169 **Q_1**	**18**	**37**	**45** **(contiene il 25%)**
170 – 174 **Q_2** **Q_3**	**25**	**62**	**76** **(contiene il 50%)** **(contiene il 75%)**
175 – 179	13	75	91
180 - 184	7	82	100
totale	82		



Rappresentazioni grafiche dei caratteri su scala numerica:

«Box plot»: una rappresentazione **sintetica** della distribuzione



Il box plot o *diagramma a scatola e baffi*, è un grafico ottenuto a partire dai 5 numeri di sintesi :

- «Minimo»
- 1° quartile (Q1)
- Mediana
- 3° quartile (Q3)
- «Massimo»

che descrivono le caratteristiche salienti della distribuzione.

N.B: Il box plot offre **una rappresentazione univoca** della distribuzione, a differenza dell'istogramma che può offrire rappresentazioni grafiche diverse a seconda degli estremi delle classi scelte.

I baffi : Q3 + 1.5*(IQR); Q1 – 1.5*(IQR) [IQR= Q3-Q1.] Dovrebbero includere il 99.3% dei dati, se provenissero da una distribuzione **gaussiana…**

Gli indici di posizione delle distribuzioni (IV): Media aritmetica

Per i caratteri quantitativi è possibile calcolare anche un altro indice di posizione: **la media aritmetica**.
Rilevate su *n* unità di una popolazione le modalità di un certo carattere quantitativo X:
$x_1, x_2, \ldots, x_n$

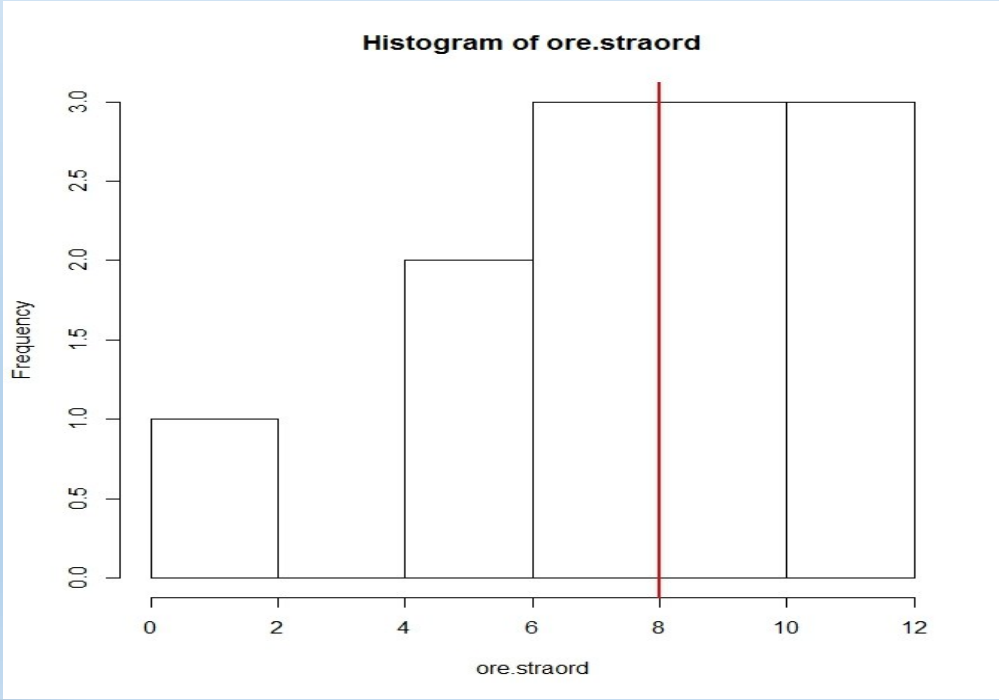
$$\bar{x} = \frac{(x_1 + x_2 + \ldots + x_n)}{n} \quad \Longleftrightarrow \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Un infermiere ha fatto delle ore di lavoro straordinario da gennaio a dicembre:

10, 12, 11, 5, 7, 10, 5, 0, 7, 10, 7, 12.

Qual è il numero medio mensile di ore di straordinario?

$$\bar{x} = \frac{(10+12+11+5+7+10+5+0+7+12)}{12} = \frac{96}{12} = 8$$



Calcolo della media aritmetica da una distribuzione di frequenze

Distribuzioni di frequenza

Variabile x	Frequenze assolute	Frequenze cumulate	Frequenze relative	Frequenze %
x_1	n_1	n_1	n_1/N	$n_1/N*100$
x_2	n_2	n_1+n_2	n_2/N	$n_2/N*100$
…	…	…	…	…
x_k	n_k	$n_1+ \ldots.+n_k=N$	n_k/N	$n_k/N*100$
totale	N		1	100

$$M = \frac{x_1 * n_1 + \cdots x_k * n_k}{N} = \frac{\sum_{i=1}^{k} x_i * n_i}{N}$$

Variabile x	Frequenze assolute
20	5
21	2
totale	7

$$M = \frac{20 \cdot 5 + 21 \cdot 2}{7}$$

M= 20.29

Calcolo della media aritmetica da una distribuzione in classi di frequenza

Valori centrali delle classi	Frequenze assolute
c_1	n_1
c_2	n_2
...	...
c_k	n_k
totale	N

$$M = \frac{c_1 * n_1 + \cdots c_k * n_k}{N} = \frac{\sum_{i=1}^{k} c_i * n_i}{N}$$

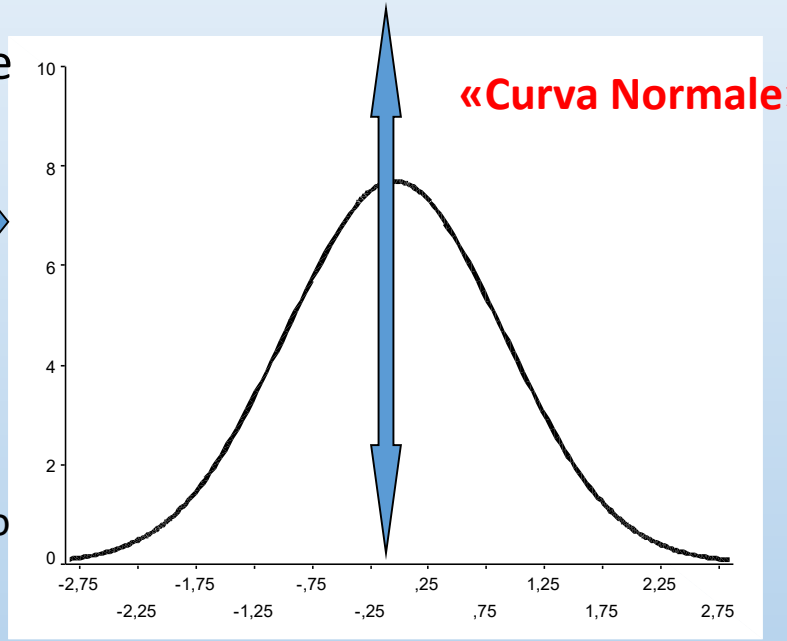
classi	Frequenze assolute	Valori centrali delle classi
[18; 22]	20	20
(22; 26]	30	24.5
(26; 30]	50	28.5
totale	100	

$$M = \frac{20 * 20 + 24.5 * 30 + 28.5 * 50}{100}$$

M=25.6

Moda, **mediana**, **quartili** e **media** sono gli indici di posizione di piu' frequente impiego.
Distribuzioni simmetriche unimodali : moda=mediana=media.
'**Simmetrico**' : una distribuzione simmetrica rispetto al valor medio: ha (circa) la stessa quantità di osservazioni inferiori alla media e superiori alla media:

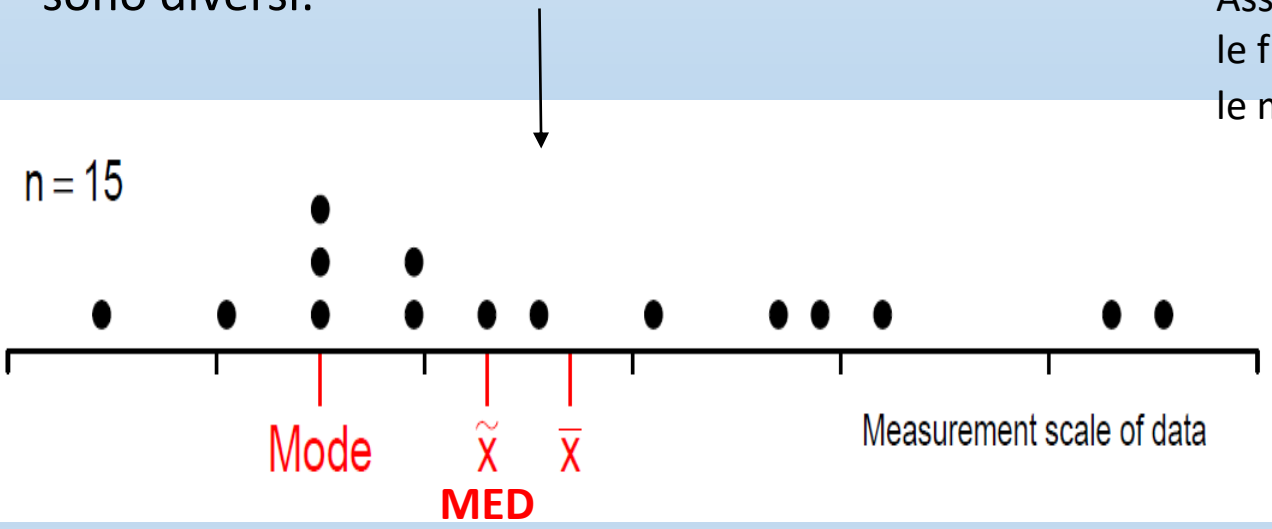
Ex: distribuzione simmetrica e unimodale
Moda=Mediana=Media



Asse verticale (y):
le frequenze con cui
le modalità si presentano

asse orizzontale (x): i valori della distribuzione
(cioè le modalità del carattere quantitativo in studio)

In caso di **distribuzione asimmetrica** invece i tre indici sono diversi:



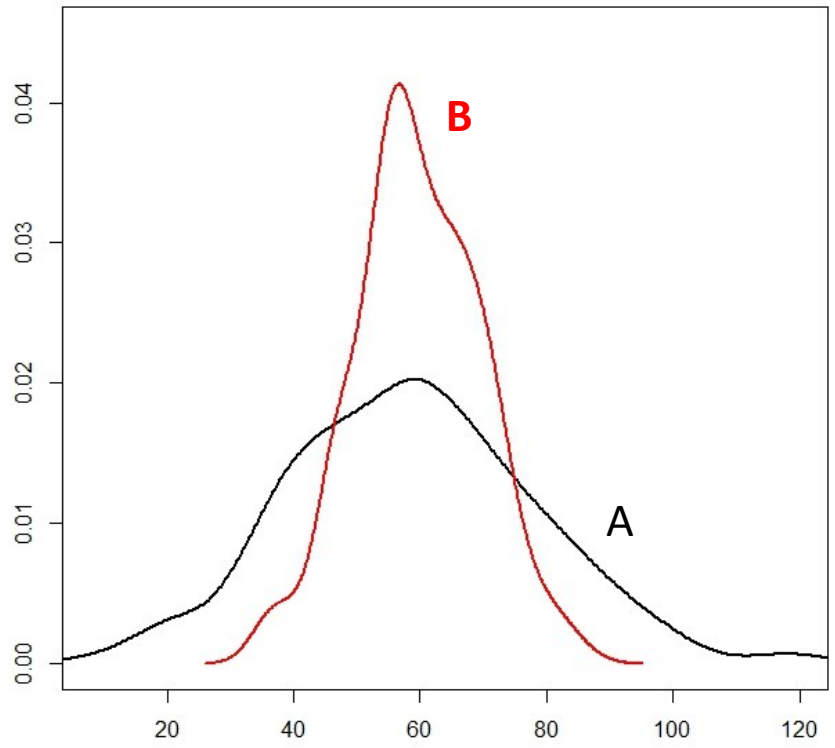
Gli indici di dispersione

La posizione è *rappresentativa* di un fenomeno; tuttavia da sola *non basta* per definire la distribuzione.
Occorrono criteri aggiuntivi per quantificare la **variabilità** delle misure rispetto ad un termine di riferimento.

La ***variabilità*** delle misure puo' essere valutata:

(a) in base alla loro ***oscillazione*** o ***dispersione*** rispetto, per esempio, al valore medio;

(b) come **distanza** tra due particolari modalità della distribuzione.



…A e B non sembrano simili…

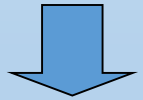
(I): Dispersione intorno al valor medio

Variabilità intorno al valore medio di una distribuzione di **valori quantitativi**: $x_1, \ldots x_n$ di media $\bar{x}$

consideriamo gli *scarti* (cioè le differenze) di tutte le misure dalla media: $\sum_{i=1}^{n} (x_i - \bar{x})$

Per proprietà della media aritmetica, la sommatoria degli scarti dalla media è sempre nulla:

(a causa della compensazione tra scarti positivi e scarti negativi). $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$



Si definisce '**devianza**' la somma dei quadrati degli scarti dalla media: $DEV = \sum_{i=1}^{n} (x_i - \bar{x})^2$

(I): Dispersione intorno al valor medio

La devianza non contiene pero' l'informazione del numero di osservazioni utilizzate nel calcolo.

$$VARIANZA = s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} = \frac{DEVIANZA}{N-1}$$

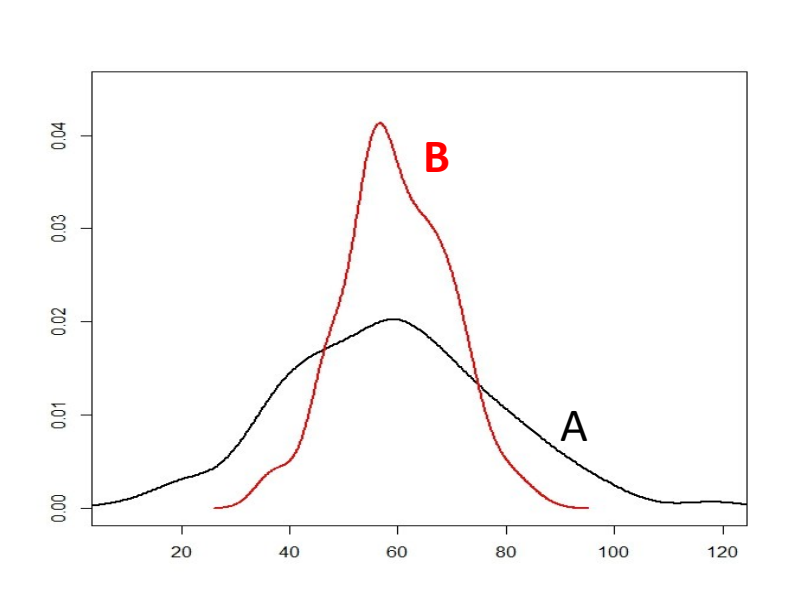
Le varianze diventano cosi' *confrontabili* tra diverse distribuzioni e si puo' stabilire quale, tra le due o piu' serie di misure considerate, presenta una maggiore dispersione rispetto alla media, **indipendentemente da N**.

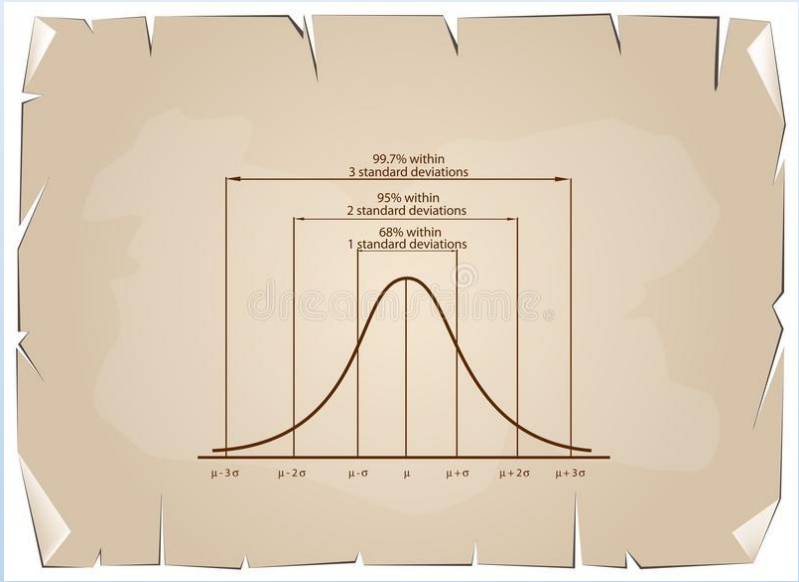
$$dev\ st = s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

Per tornare all'unità di misura del fenomeno si usa *la radice quadrata* => deviazione standard

(I): Dispersione intorno al valor medio

Deviazione standard: errore che si commette *mediamente* considerando il valore medio al posto di ogni singolo valore della distribuzione.





E' **<u>MOLTO IMPORTANTE</u>** calcolare la deviazione standard dalla media perchè è un indice fondamentale per capire se i dati che stiamo osservando siano ***ben sintetizzati*** dalla media oppure no.

Quanto piu' è alta infatti la deviazione standard, tanto meno è informativa la media, perchè i dati si allontanano molto da essa.

Excursus: perché N-1 ?

In *piccoli* studi si riduce la probabilità di includere dati particolarmente "*dispersi*" e quindi si può *sottostimare* la variabilità reale della popolazione.

Queste considerazioni giustificano, per N «piccolo» (<50) di modificare la formula della varianza:

$$VARIANZA = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N-1}$$

Motivo statistico-matematico: se di una distribuzione di dati è nota la media, la conoscenza di uno degli N valori è superflua in quanto ricavabile utilizzando la formula del calcolo della media stessa.

Ne deriva che questa misura, pur essendo indispensabile per determinare la media, non è '**indipendente**' dalle altre, in quanto è implicita in esse e non fornisce ulteriore informazione; per tale motivo viene trascurata nel calcolo della varianza.

Le N-1 osservazioni *indipendenti*, costituiscono un caso particolare di ***gradi di libertà***: un concetto statistico importante, soprattutto in inferenza, per indicare, nelle varie situazioni, il **numero delle *informazioni indipendenti***.

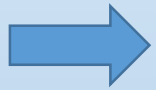
Esempio di calcolo della deviazione standard (distribuzione di frequenze)

Variabile x	Frequenze assolute
x_1	n_1
x_2	n_2
…	…
x_k	n_k
totale	N

$$dev\ st = s = \sqrt{\frac{\sum_{i=1}^{k} n_i * (x_i - M)^2}{N-1}}$$

Variabile x	Frequenze assolute
20	3
22	4
26	5
totale	12

$$M = \frac{20 * 3 + 22 * 4 + 26 * 5}{12}$$



M=23.17

Variabile X	Frequenze assolute	(x-M)	(x-M)**2	freq*(x-m)**2
20	3	-3,17	10,05	30,15
22	4	-1,17	1,37	5,48
26	5	2,83	8,01	40,04
Totale	12		Varianza:	6,88
			Dev std:	**2,62**

Gli indici di dispersione (II): distanza

La dispersione come **distanza** tra due particolari modalità della distribuzione:

a) Distanza tra il valore minimo ed il valore massimo

b) Distanza tra il *primo* ed il *terzo* quartile della distribuzione

a) Data una distribuzione di valori $x_1, x_2, x_3 \ldots, x_n$ di un **carattere qualitativo ordinabile** o **quantitativo**, ordinata in senso crescente: $x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}$ si definisce:

INTERVALLO DI VARIAZIONE (**RANGE**): distanza tra il valore minimo $x_{(1)}$ ed il valore massimo $x_{(n)}$ della distribuzione ordinata:

$$\text{RANGE} = x(n) - x(1)$$

b) Data una distribuzione di valori $x_1, x_2, x_3 \ldots, x_n$ di un **carattere qualitativo ordinabile** o **quantitativo**, si definisce:

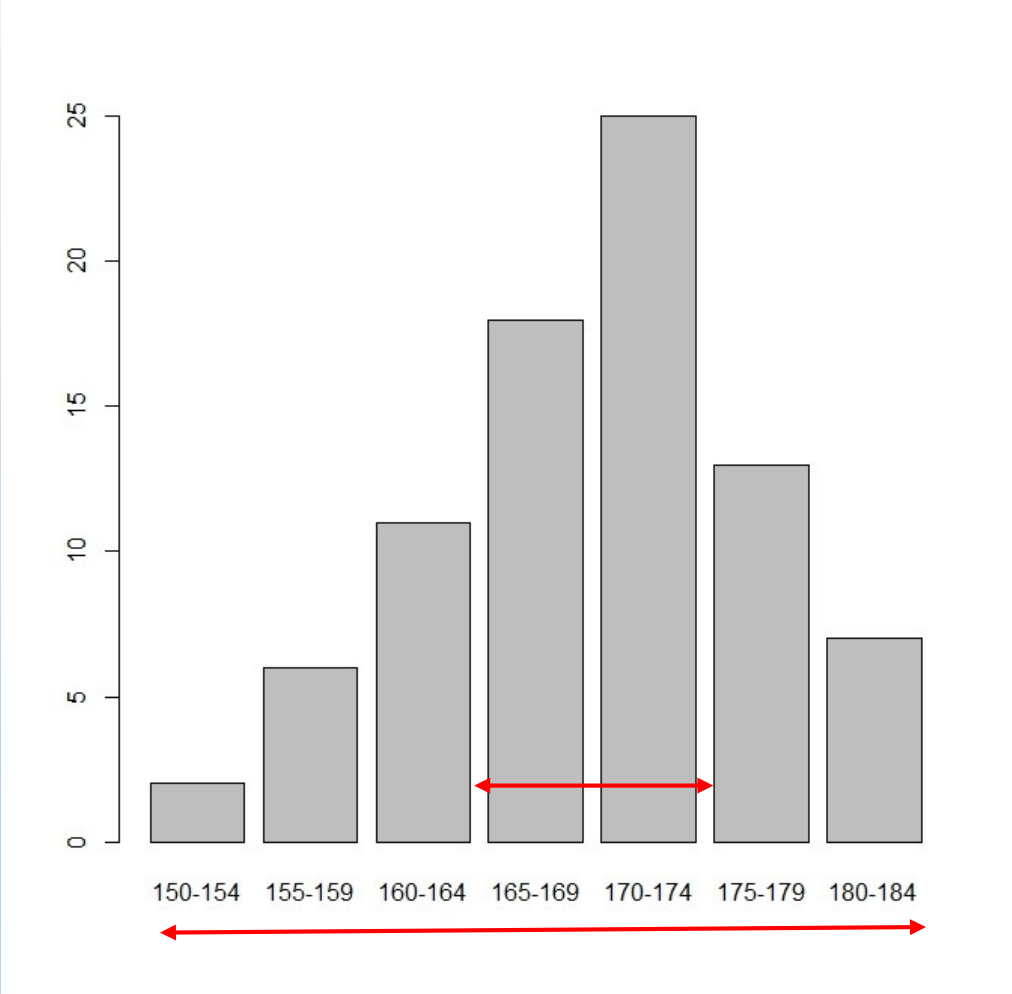
DIFFERENZA INTERQUARTILE (**RANGE INTERQUARTILE, IQR**) la distanza tra il primo ed il terzo quartile:

$$\text{RANGE INTERQUARTILE} = Q_3 - Q_1$$

Gli indici di dispersione (II): distanza

Questi indici sono detti ***ASSOLUTI*** perché sono espressi nella stessa unità di misura del carattere.

Classi di altezza (cm)	Frequenze assolute	Frequenze cumulate assolute	Frequenze cumulate percentuali (%)
150 – 154	2	2	2
155 – 159	6	8	10
160 – 164	11	19	23
165 – 169	18	37	45
170 – 174	25	62	76
175 – 179	13	75	91
180 - 184	7	82	100
totale	82		



RANGE = 184 – 150 = 34 cm

Q1=165-169 e Q3=170-174

IQR= 172 – 167 = 5 cm (per convenzione si può utilizzare il valor medio delle classi)

Gli indici di dispersione (III): il coefficiente di variazione

Per confrontare la dispersione **di due o piu'** distribuzioni si può ricorrere al '**coefficiente di variazione**':

$$CV\% = \frac{s}{\bar{x}} * 100$$

CV%= Rapporto tra la deviazione standard e la media in %.

'Numero puro' (non espresso in una determinata unità di misura) e quindi confrontabile con altri.

Utile anche per fenomeni che abbiano ***una media molto diversa*** (pur avendo la stessa unità di misura).

Ex: verificare la precisione di analisi di laboratorio, chimico-cliniche:

- variabilità **intra** operatore
- **inter** operatori
- **inter/intra** laboratori

Gli indici di dispersione (III): il coefficiente di variazione

Quale tra glicemia e calcemia è piu' dispersa rispetto alla media?

$m_{glicemia}$ = 85 mg/100 ml
$s_{glicemia}$ = 11 mg/100 ml
$m_{calcemia}$ = 9 mg/100 ml
$s_{calcemia}$ = 1,5 mg/100 ml

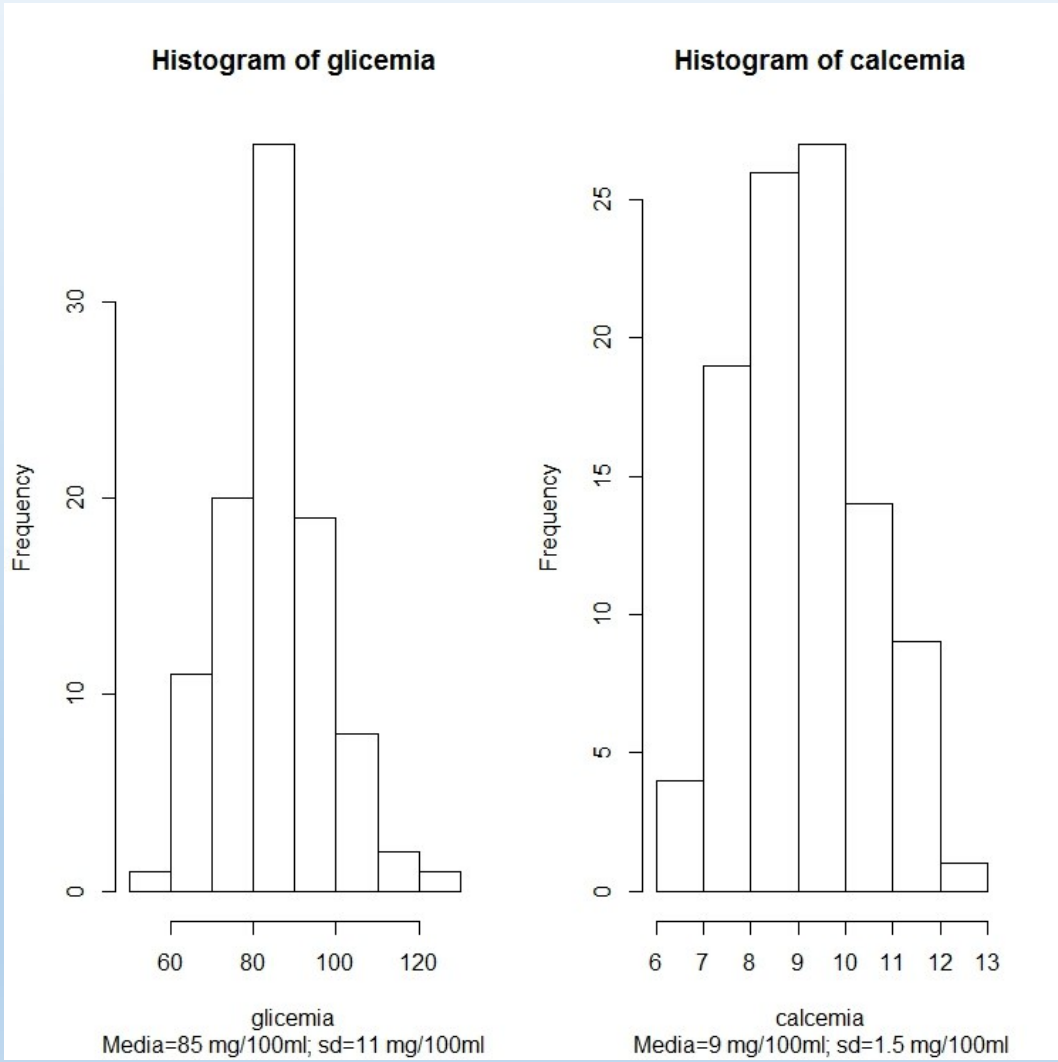
CV di Glicemia

$$\frac{11\,mg/100\,ml}{85\,mg/100\,ml} * 100 = 12.9\%$$

CV di Calcemia

$$\frac{1.5\,mg/100\,ml}{9\,mg/100\,ml} * 100 = 16.7\%$$

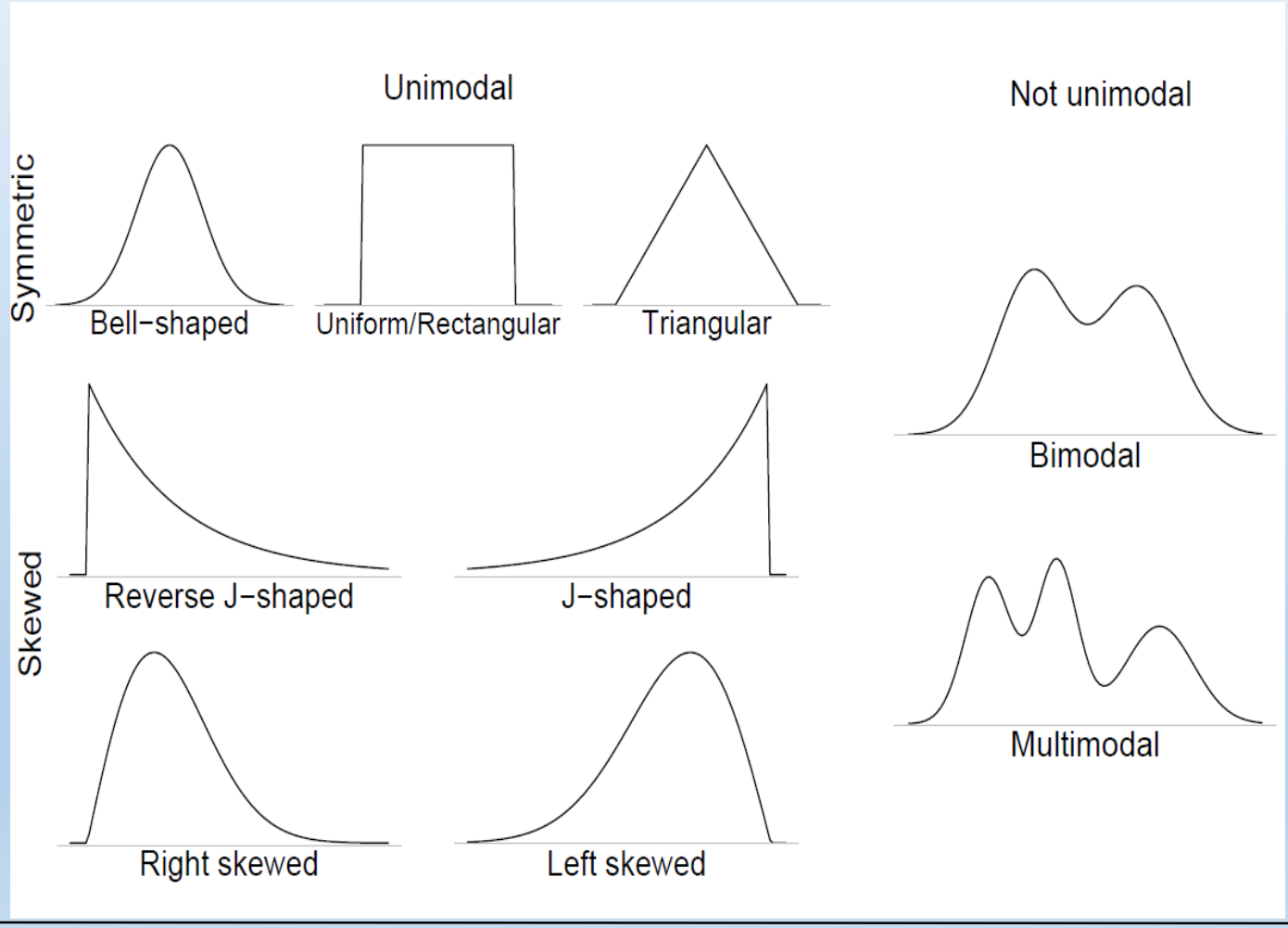
La calcemia ha un grado di dispersione maggiore della glicemia.



E' necessario visualizzare i dati per scegliere un opportuno indice di posizione.
Per distribuzioni asimmetriche è preferibile usare la MEDIANA perché la MEDIA risente eccessivamente di valori *anomali* (estremamente grandi o estremamente piccoli).

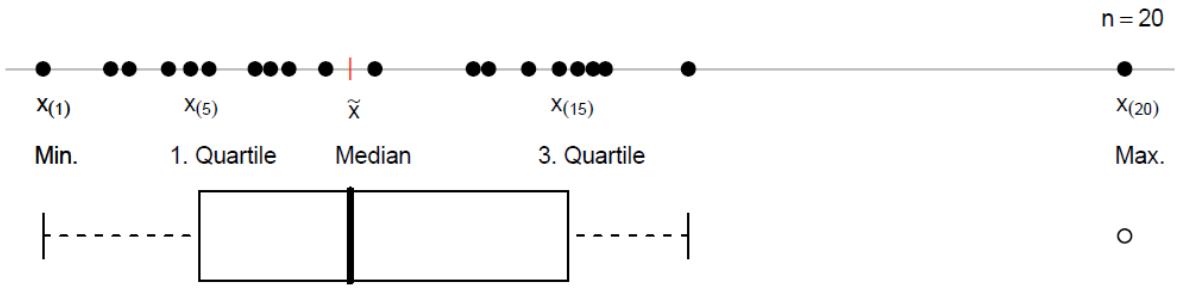
Quali indici di posizione sono calcolabili per i vari tipi di caratteri:

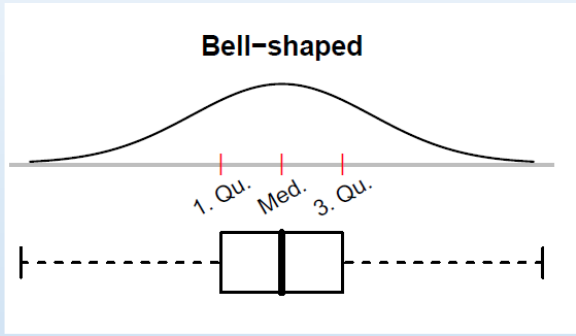
CARATTERE		INDICE DI POSIZIONE
qualitativo	Sconnesso	Moda
	Ordinabile	Moda, mediana, quartili
quantitativo		Moda, mediana, quartili, media

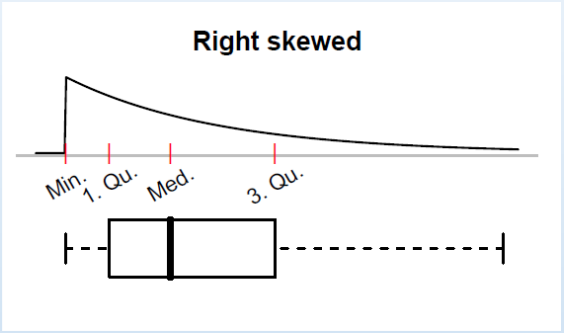


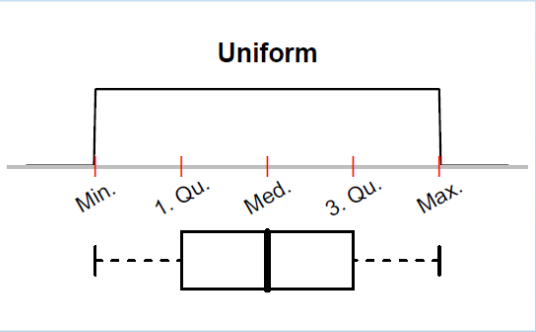
E' **importante** ricordare l'associazione tra il tipo di carattere ed i relativi indici di dispersione che possono essere calcolati:

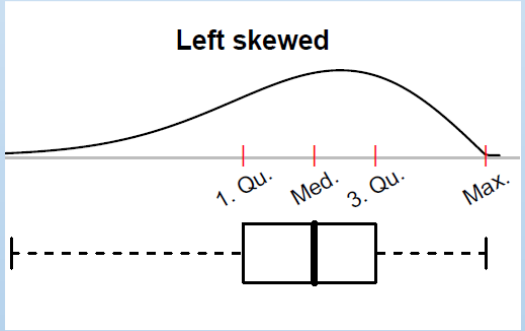
CARATTERE		INDICE DI DISPERSIONE
qualitativo	Sconnesso	nessuno
	Ordinabile	RANGE / IQR *[modalità inferiore/superiore oppure modalità relative a Q1 e Q3 del carattere qualitativo ordinabile, non certo come differenza tra esse !!]*
quantitativo		RANGE / IQR DEVIANZA, VARIANZA, DEVIAZIONE STANDARD e CV%









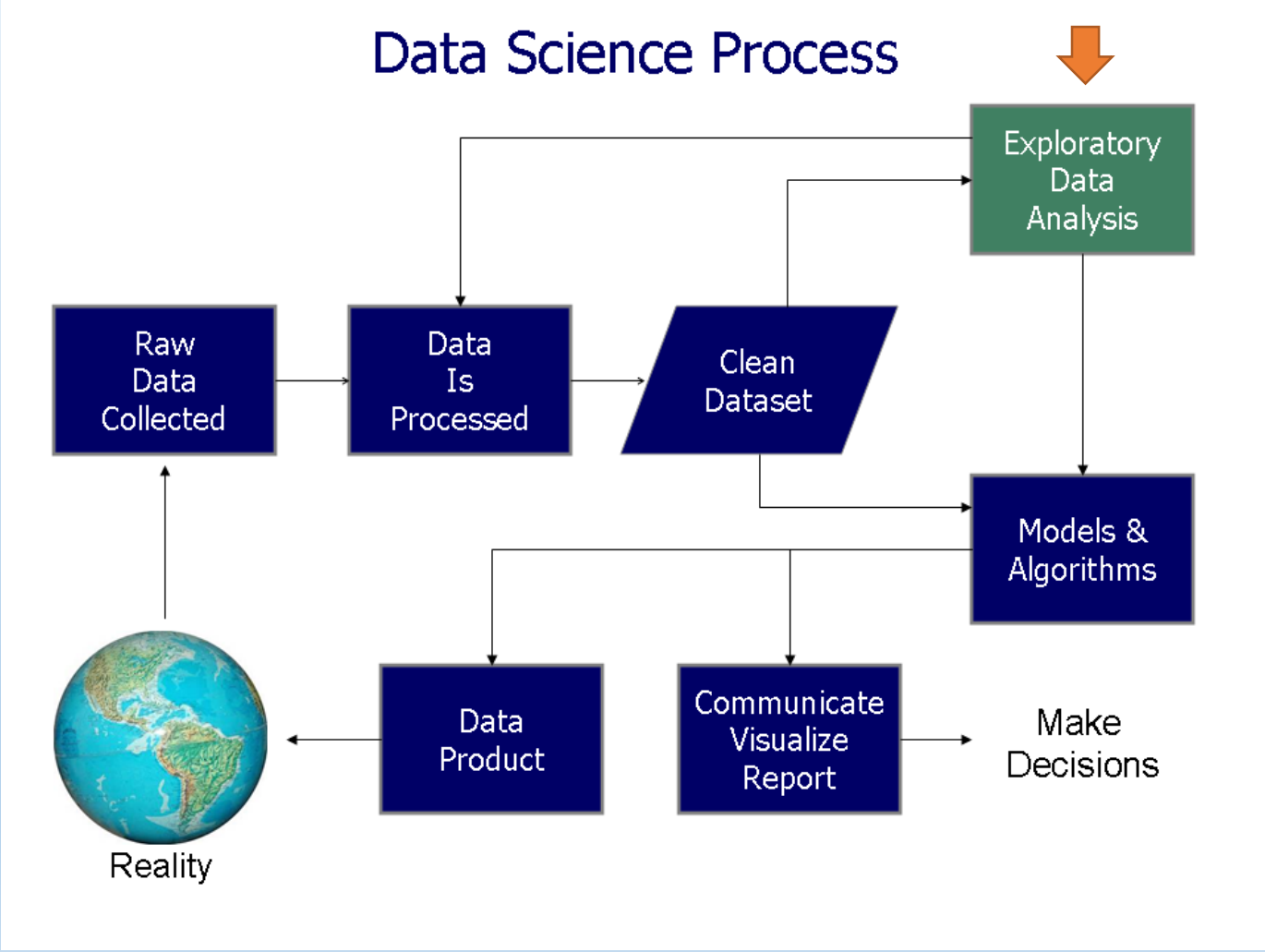


Per una corretta e completa *sintesi* dei dati all'indicazione della tendenza centrale va associata necessariamente l'informazione della loro dispersione.

Indici di Forma

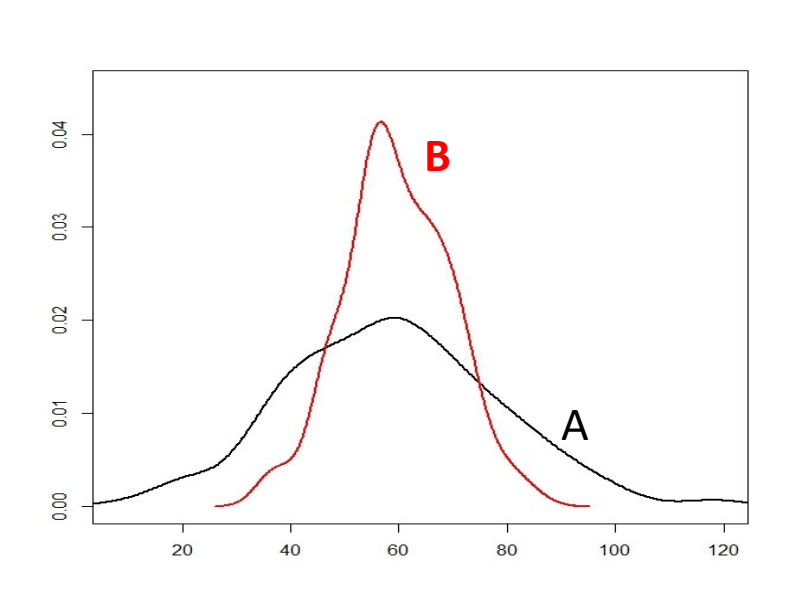
There are lies, damned lies and statistics.

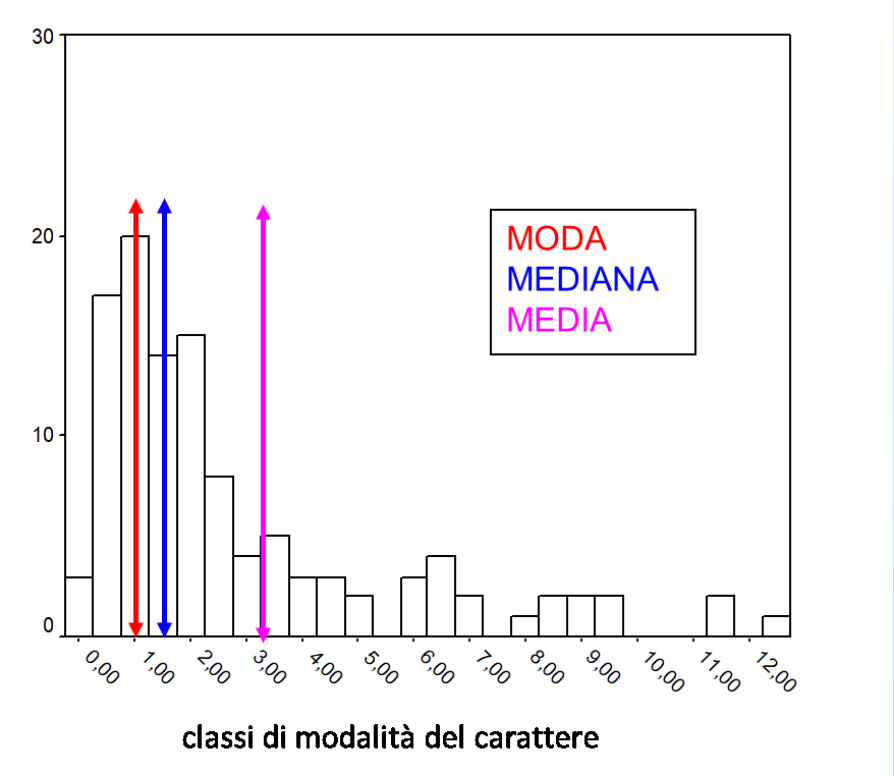
Mark Twain



Media, moda e mediana forniscono indicazioni utili sulla **forma** delle distribuzioni: le loro rispettive posizioni e le differenze tra i loro valori indicano se vi è uno **sbilanciamento** della distribuzione verso destra o verso sinistra.

Però, anche a parità di media e mediana, le distribuzioni possono avere diversa **variabilità**, presentando quindi *forme* differenti.

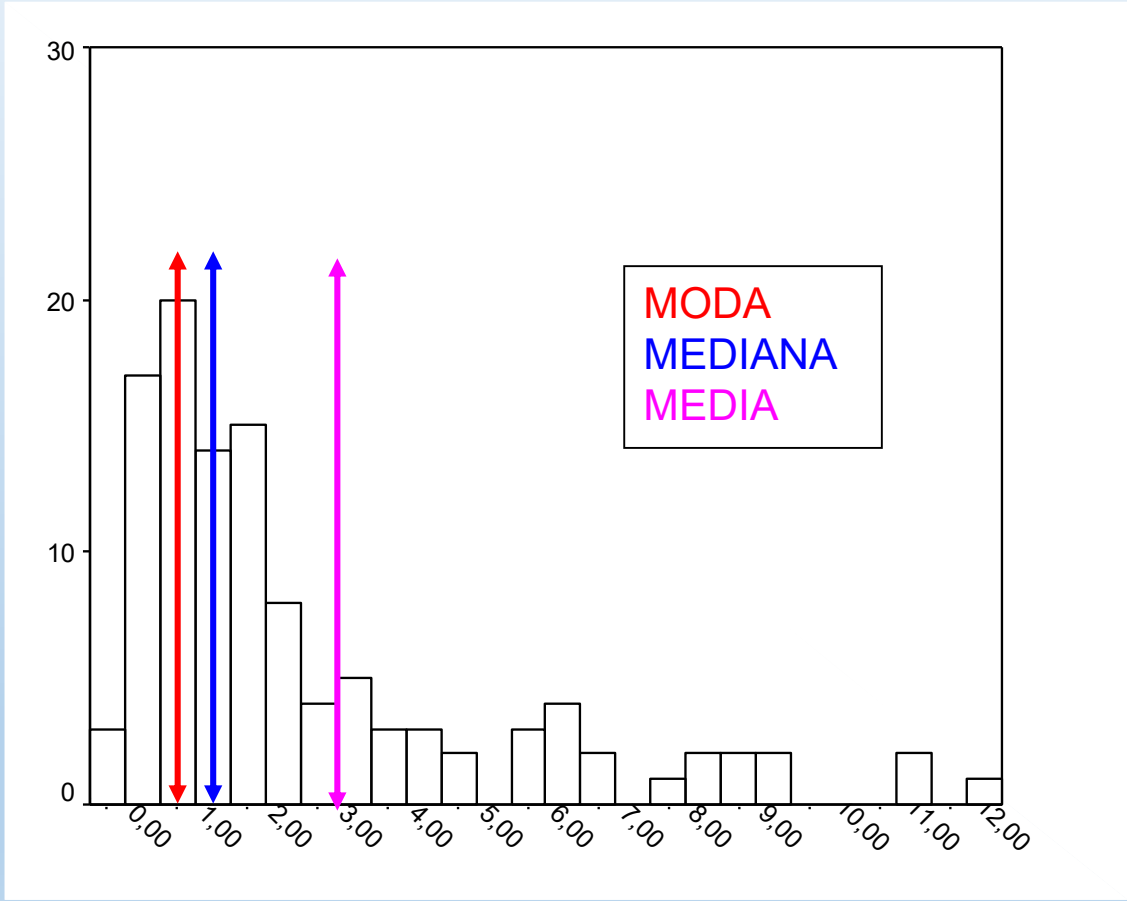




Esistono diversi indici che misurano quanto una distribuzione sia «appiattita» o «appuntita» (si parla in questo caso di curtosi) e l'entità degli eventuali sbilanciamenti verso destra o sinistra (asimmetria).

Gli indici di «forma» delle distribuzioni (I): asimmetria (skewness)

Quanto piu' moda, mediana e media si differenziano,
tanto piu' la distribuzione è asimmetrica.



classi di modalità del carattere

Indice di **skewness** (=asimmetria):

$$\frac{1}{n}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3=\frac{1}{n}\sum z_i^3 \longrightarrow z_i=\frac{x_i-\bar{x}}{s}$$

'z-score' di x_i

La distribuzione è
simmetrica se Skewness=0 (circa)

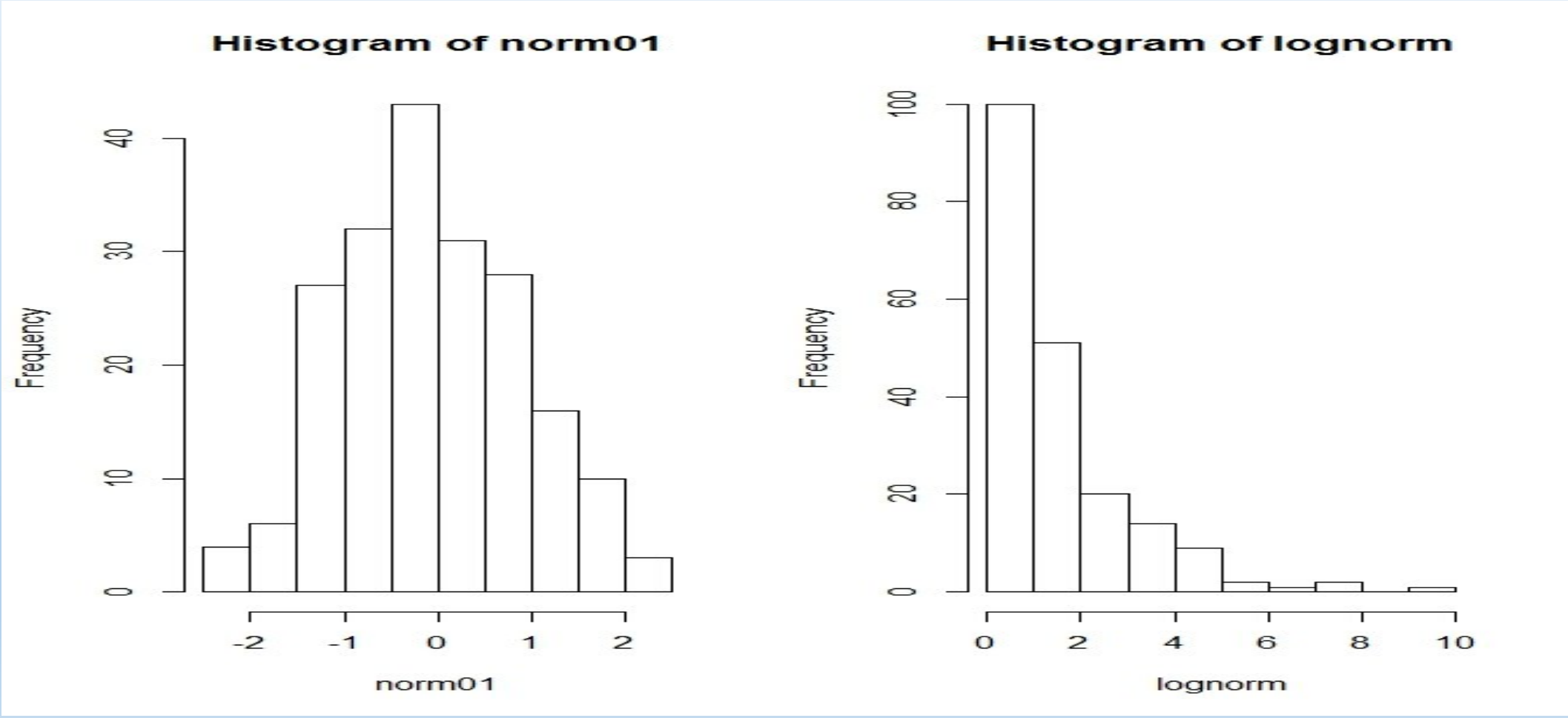
è asimmetrica a destra (positiva) se Skewness > 0

è asimmetrica a sinistra (negativa) se Skewness < 0

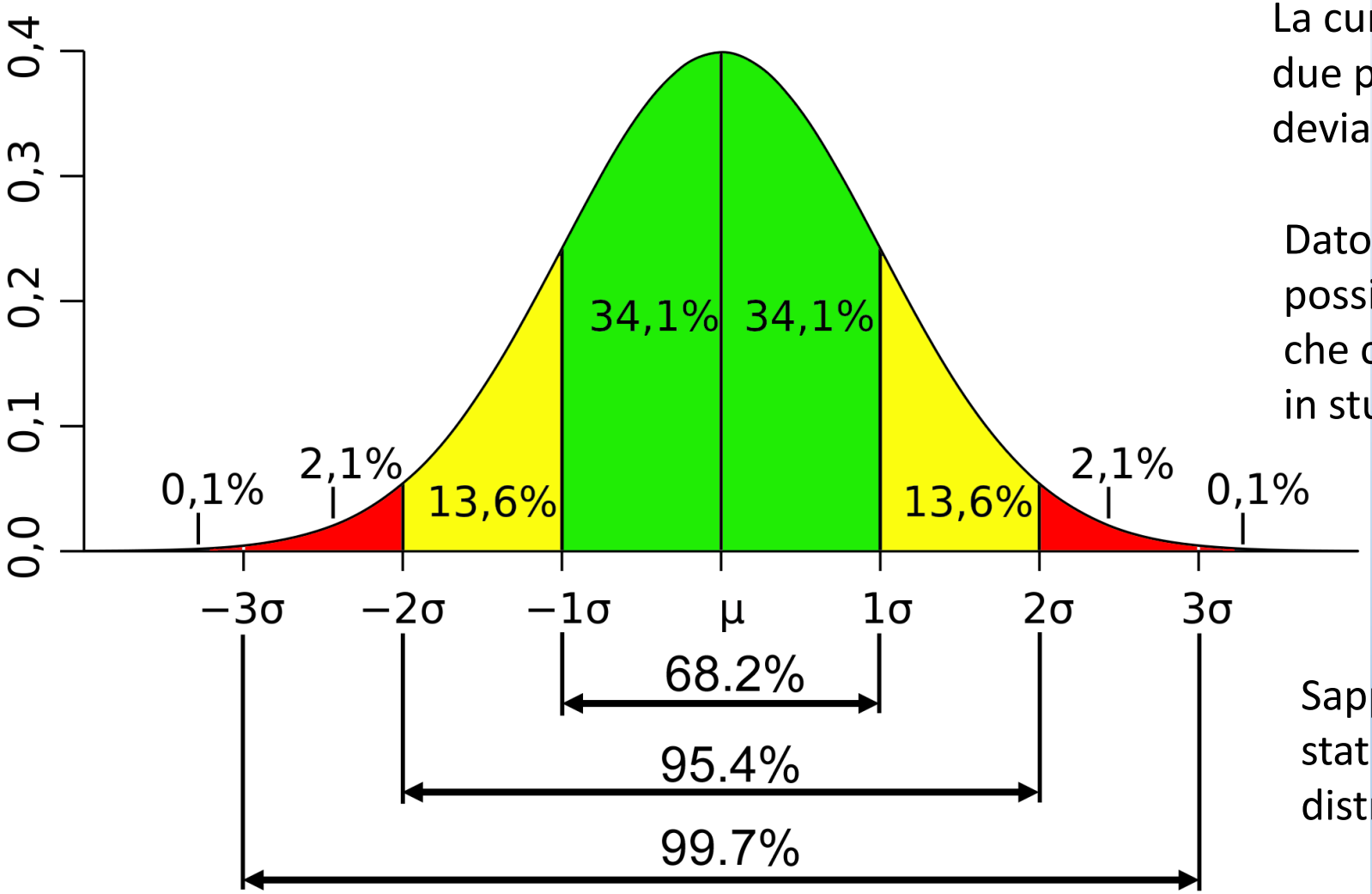
Gli indici di «forma» delle distribuzioni (I): asimmetria (skewness)

```
skewness(norm01)     # 0.17
```

```
skewness(lognorm)    # 2.11
```



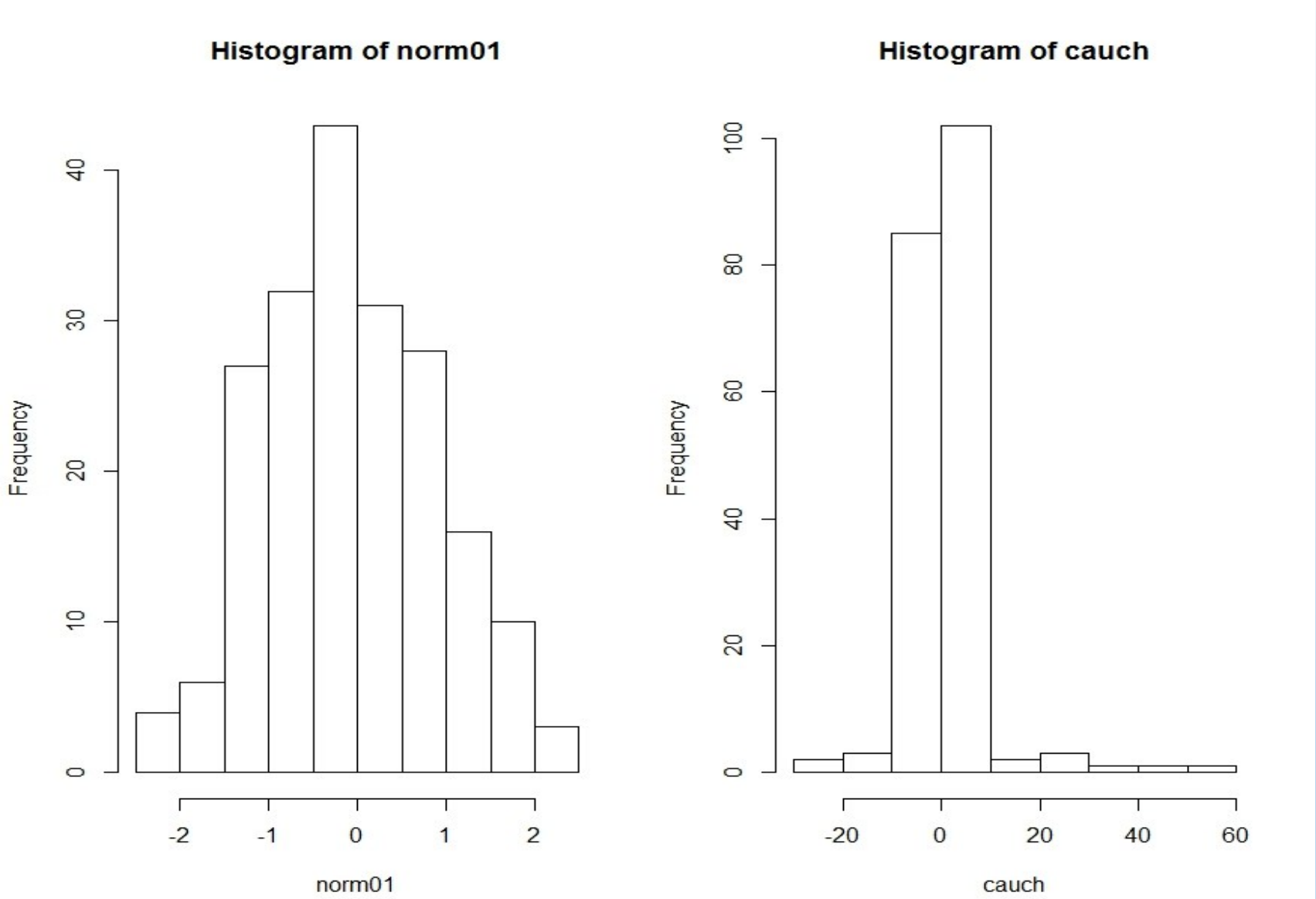
La «forma» della distribuzione normale (curva gaussiana)



La curva normale è caratterizzata da due parametri: la media μ e la deviazione standard σ.

Dato il valore di questi due parametri, possiamo calcolare la % di unità statistiche che cadono in certe regioni della variabile in studio.

Sappiamo quindi anche la % di unità statistiche che cadono nelle «code» della distribuzione.



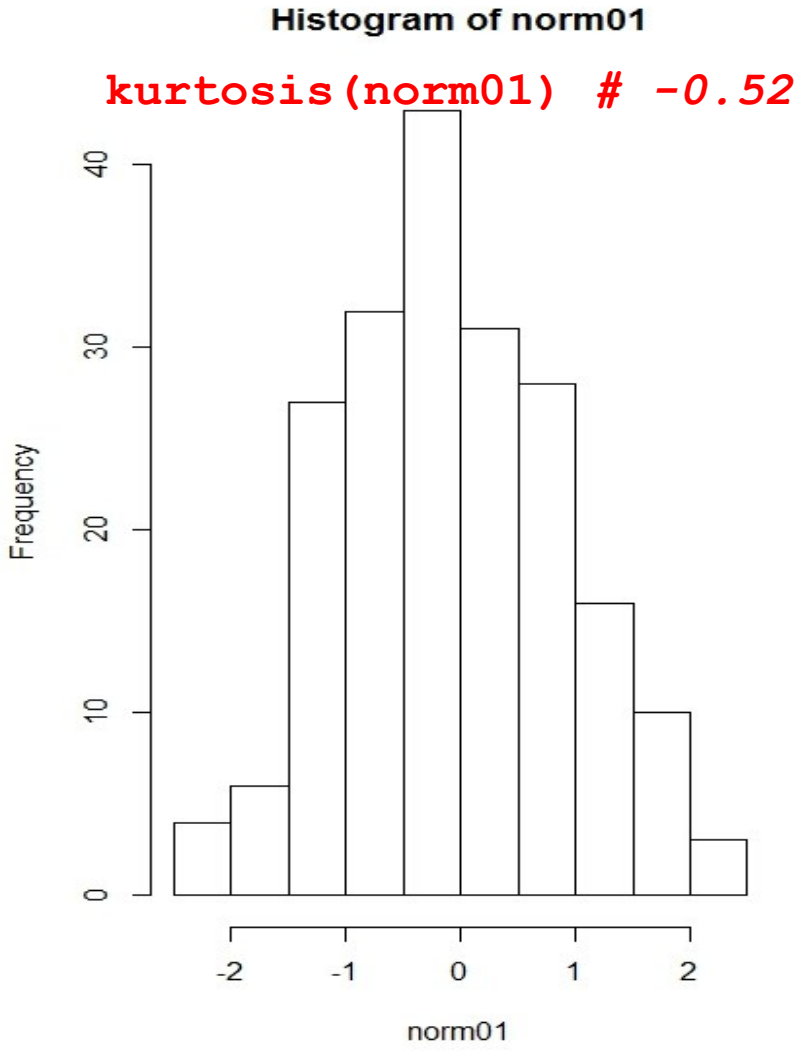
La **curtosi** è un indice che valuta la «pesantezza» delle code relativa al resto della distribuzione (cioè la % di unità statistiche presenti nelle code della distribuzione rispetto al totale).

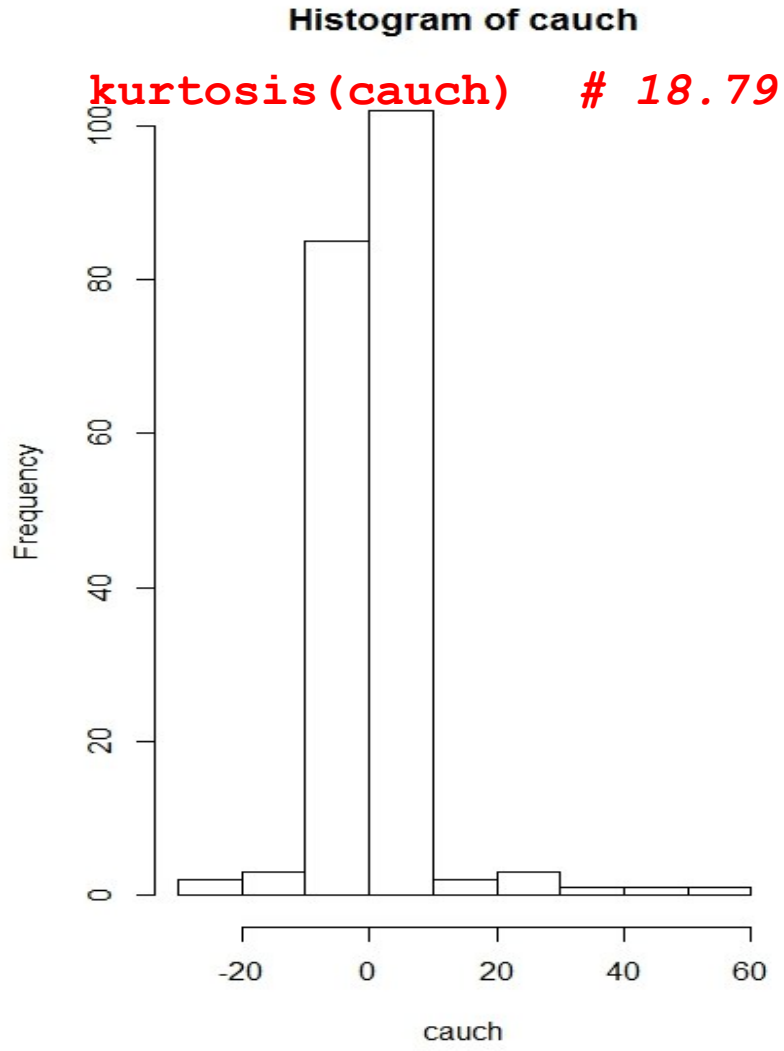
L' indice viene calcolato rispetto alla % teorica che dovrebbe essere presente se la distribuzione fosse normale (con uguale media e varianza).

Gli indici di «forma» delle distribuzioni (II): curtosi

$$\frac{\sum \frac{(x_i - \bar{x})^4}{\sigma^4}}{n} - 3 = \frac{1}{n}\sum z_i^4 - 3$$

Indice di **curtosi**: «3» è il valore di curtosi della curva gaussiana

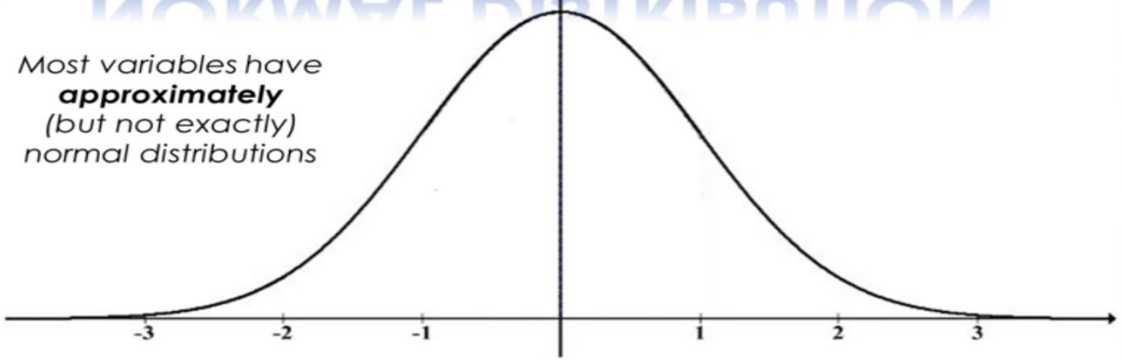




*Il calcolo del coefficiente di curtosi ha senso solo nelle distribuzioni monomodali

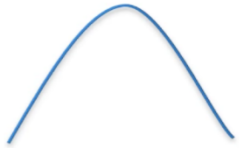
NORMAL DISTRIBUTION

Most variables have **approximately** (but not exactly) normal distributions

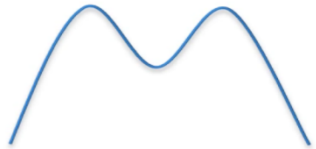


Modality

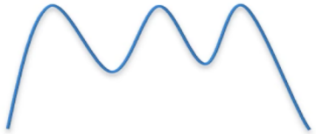
Unimodal
one-peak



Bimodal
two-peaks

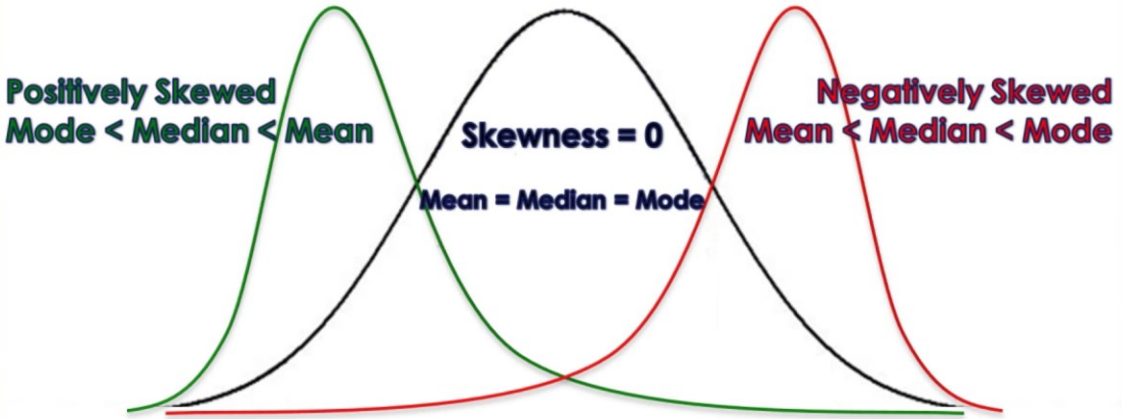


Multimodal
two or more peaks

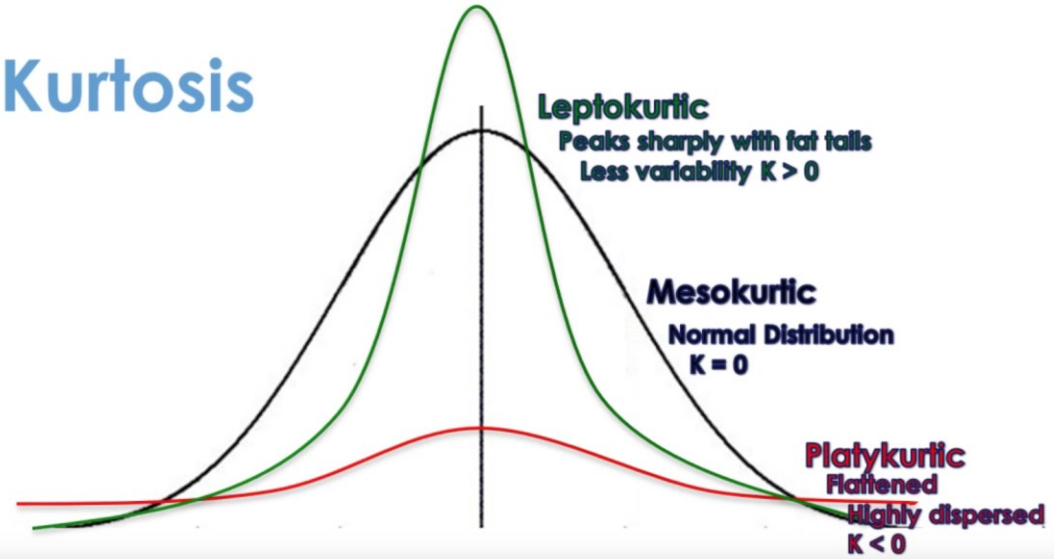


Skewness

$$Skewness = \frac{(Mean - Median)}{Standard\ Deviation}$$



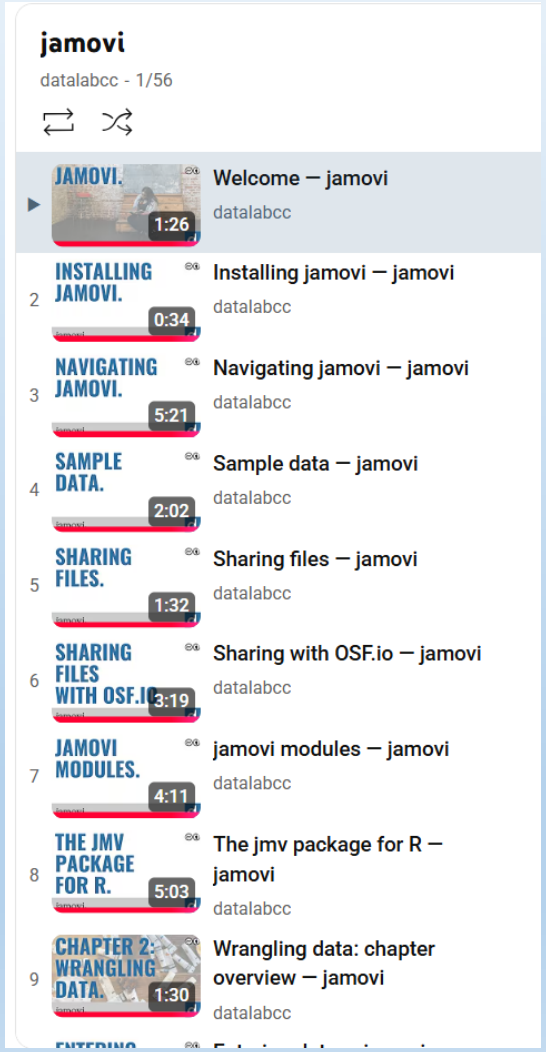
Kurtosis

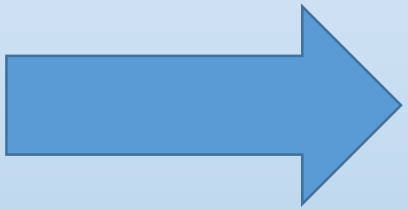


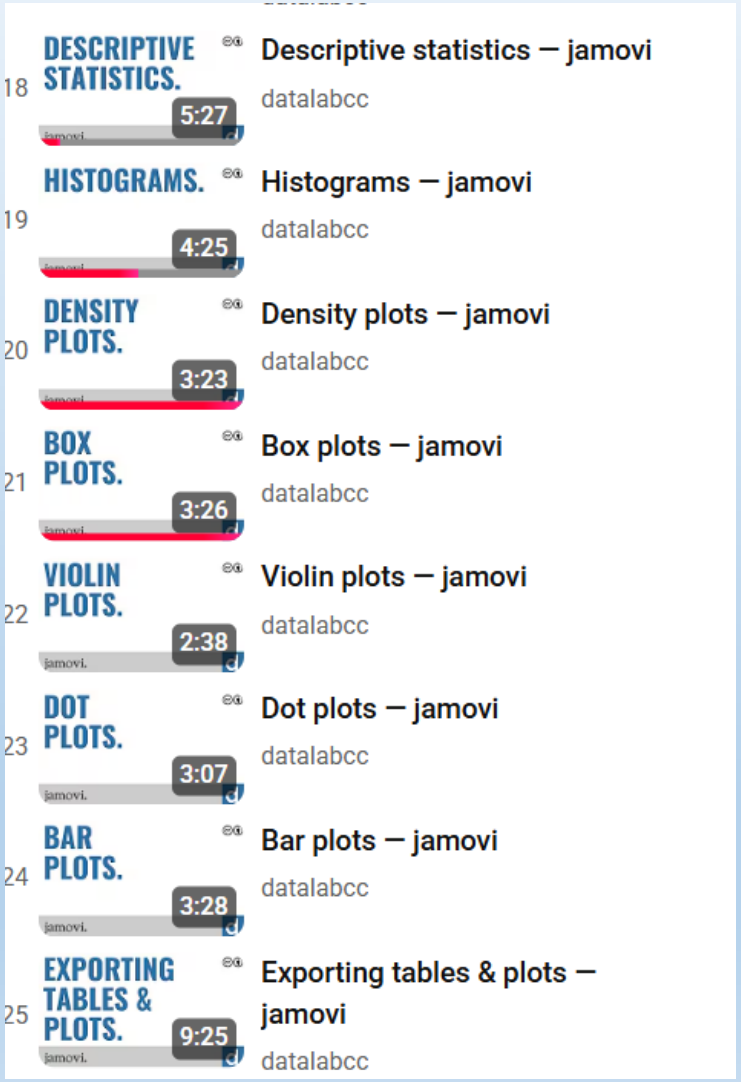
Divertitevi ad utilizzare JAMOVI !

https://datalab.cc/jamovi/

https://www.youtube.com/watch?v=Ej9e8lzaeDE&list=PLkk92zzyru5OAtc_ItUubaSSq6S_TGfRn







Supplementary materials

Trasformazioni delle scale [cenni!]

La **trasformazione di scala** di una variabile (su scala quantitativa) può cambiare la sua distribuzione da asimmetrica a una distribuzione più simmetrica, più *simile* alla distribuzione normale.

Questo passaggio è utile per vari scopi: molti metodi statistici (inferenziali) assumono la normalità delle variabili in esame.

Inoltre, se stiamo studiando l'associazione tra una coppia di variabili, opportune trasformazioni della loro scala possono migliorare la *linearità* della associazione.

Inoltre, se stiamo confrontando più distribuzioni, una volta soddisfatta la «*normalità*», possiamo procedere al calcolo dei loro **valori standardizzati**, che possono risultare utili in molte procedure.

Bisogna però poi fare attenzione a come vengono riportati i risultati delle analisi con le variabili trasformate.

Per presentare i valori degli indici di posizione o dispersione, è a volte consigliabile *ri-trasformare* sulla scala originale per evitare ambiguità nell'interpretazione.

Standardizzazione delle variabili (su scala quantitativa)

In statistica, per "**standardizzazione**" si intende la trasformazione di una variabile quantitativa per renderla più facilmente confrontabile con le altre.

Una variabile standardizzata è una variabile quantitativa a cui è stata cambiata la scala di misurazione ottenendo dei *numeri puri* (detti anche **punteggi z** o punteggi standard). Questi nuovi valori sono detti anche *adimensionali*, in quanto sono svincolati dall'unità di misura della variabile di partenza.

La caratteristica principale di una variabile standardizzata è poi che ha sempre media=0 e deviazione standard=1.

La standardizzazione permette quindi di **confrontare** variabili che hanno medie e deviazioni standard misurate su diversa unità di misura o ordine di grandezza. Ad esempio, per capire se c'è più variabilità tra il peso (in kg) o l'altezza (in cm) di un gruppo di individui.

Come si ottiene una variabile standardizzata:

- Come prima cosa di calcola la media e la deviazione standard della variabile [se la distribuzione è simmetrica…].
- Successivamente, ai singoli valori della variabile viene sottratta la media.
- Il risultato di tale differenza viene diviso per la deviazione standard della variabile stessa.

In questo modo, si ottiene una variabile standardizzata che ha sempre media=0 e deviazione standard=1.

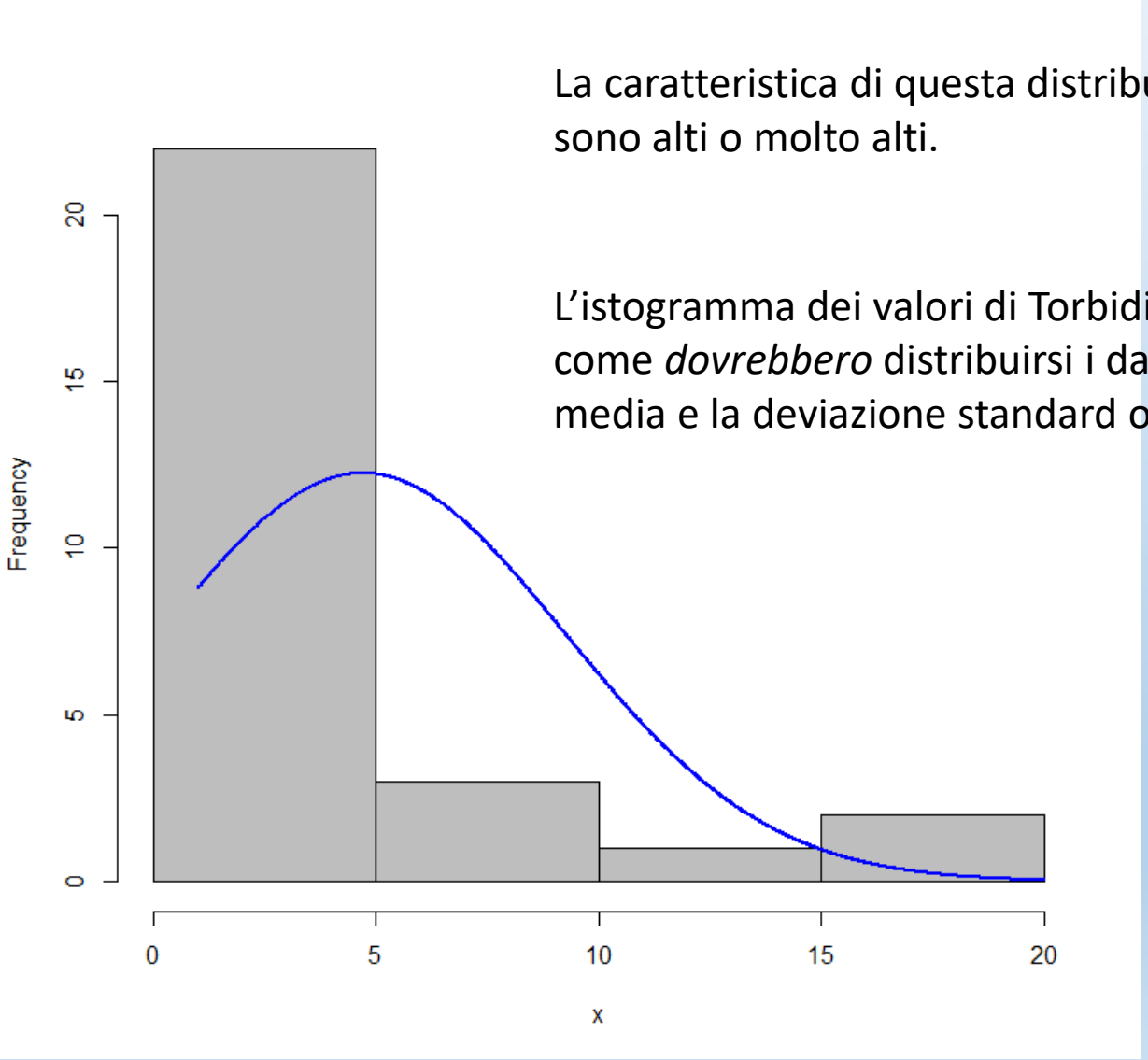
Calcolo di un punteggio standardizzato: due esempi

Ipotizziamo che un campione di individui abbia un peso medio di 70 kg ed una deviazione standard di 10 kg. Standardizzare la variabile peso significa prendere i singoli pesi degli individui e per ognuno di essi sottrarre 70 e poi dividere il risultato per 10.

Ad esempio, il valore standardizzato per Giovanni, che pesa 85 kg, sarà pari a (85-70)/10= **+1.5**. Il valore standardizzato di Maria, che invece di chili ne pesa 50, sarà (50-70)/10= -2.

TRASFORMAZIONE DELLA SCALA DI MISURA

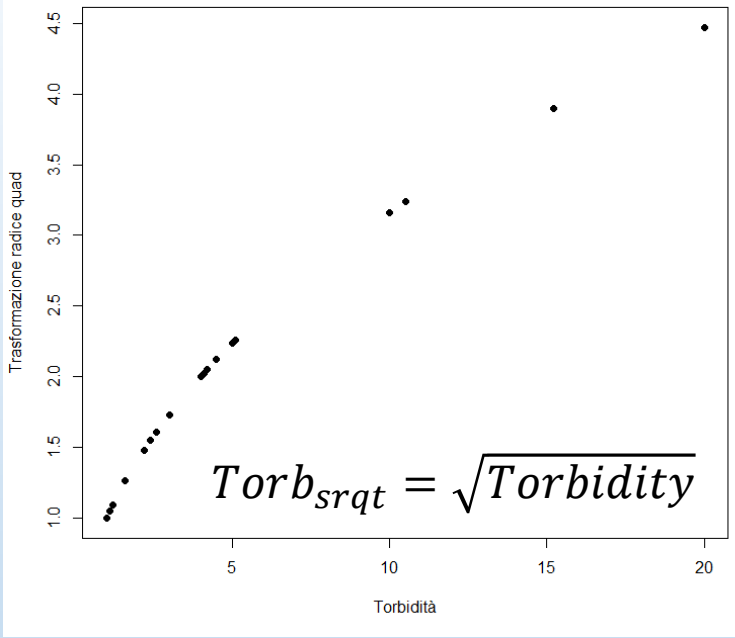
Questo esempio utilizza dei dati ipotetici della torbidità (x) dell'acqua del fiume.
La torbidità è una misura di quanto l'acqua possa diventare torbida a causa del materiale sospeso (inquinanti per lo più).

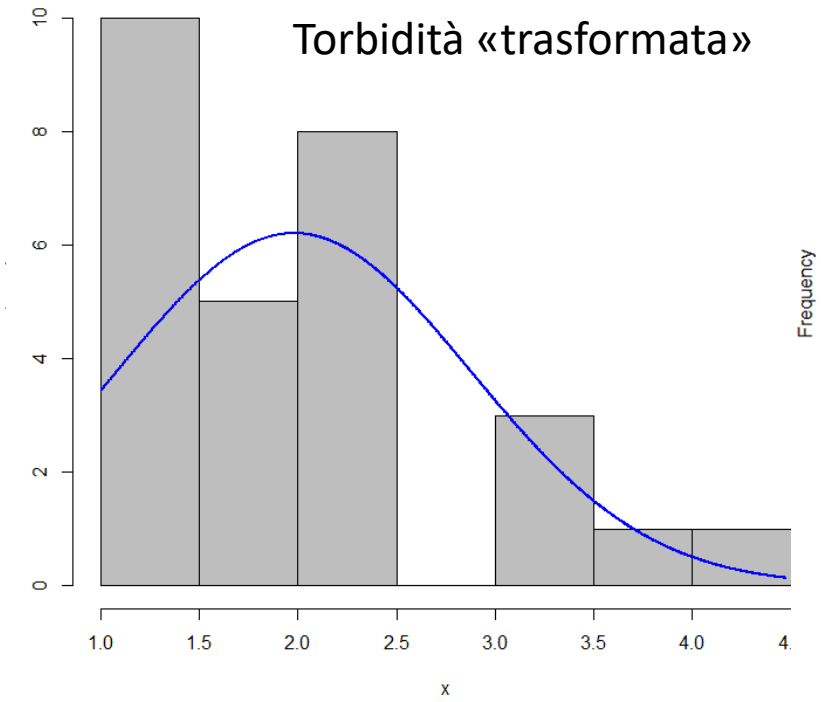


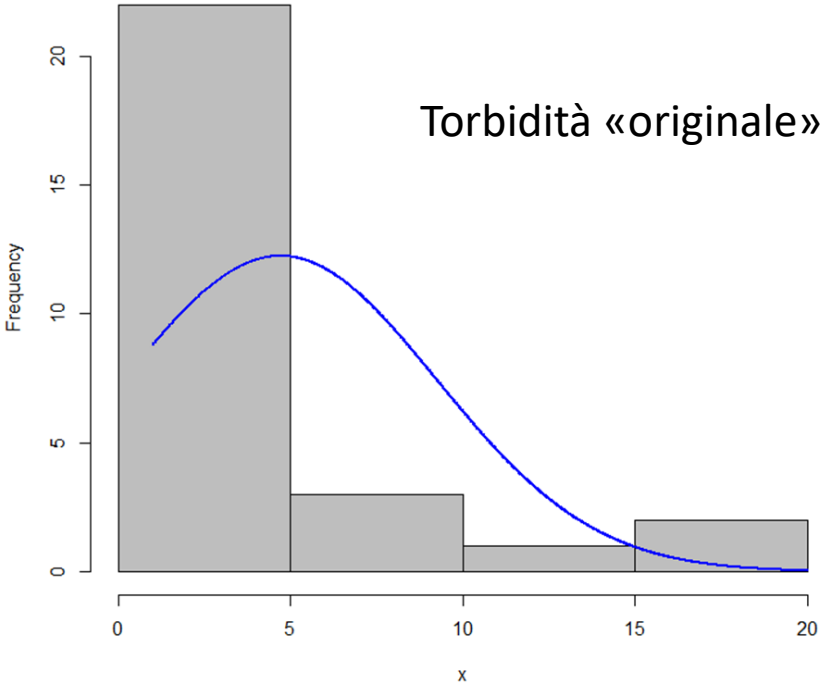
La caratteristica di questa distribuzione è che i valori sono per lo più bassi, ma occasionalmente sono alti o molto alti.

L'istogramma dei valori di Torbidità, con una **curva normale (blu)** sovrapposta [ovvero come *dovrebbero* distribuirsi i dati, se provenissero da una distribuzione normale, data la media e la deviazione standard osservate] ci mostra la forte asimmetria a destra.

Proviamo adesso ad applicare alcune delle trasformazioni più comuni per i dati asimmetrici: la radice quadrata, la radice cubica ed il logaritmo.



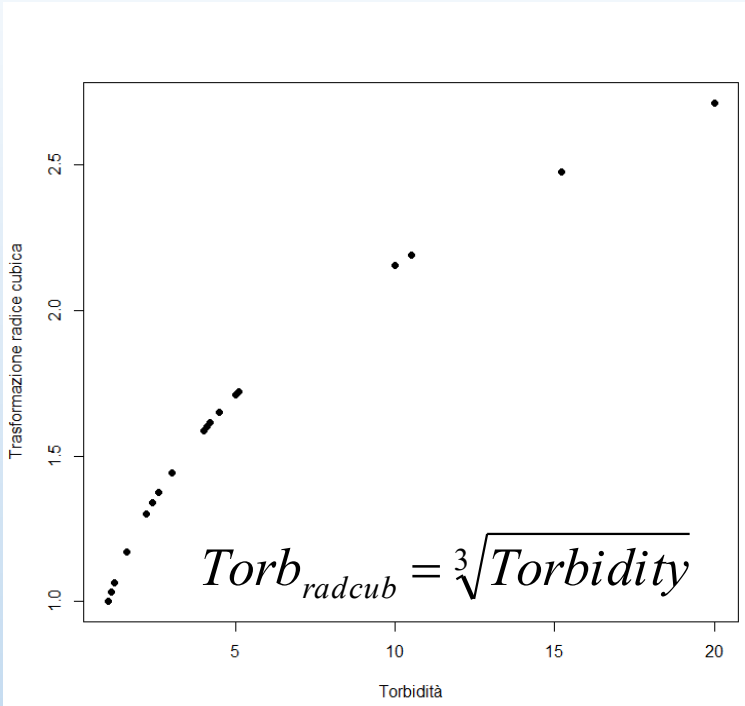


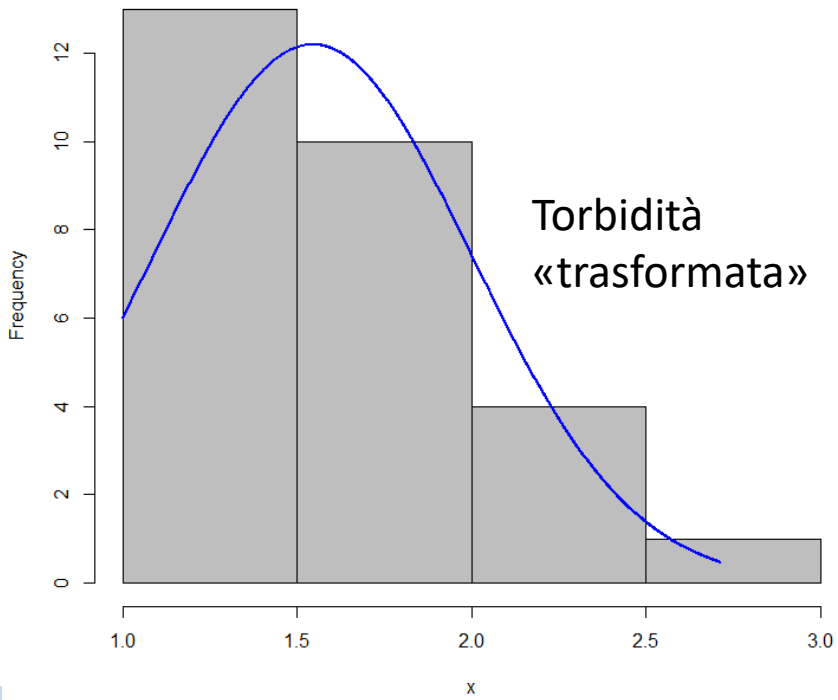


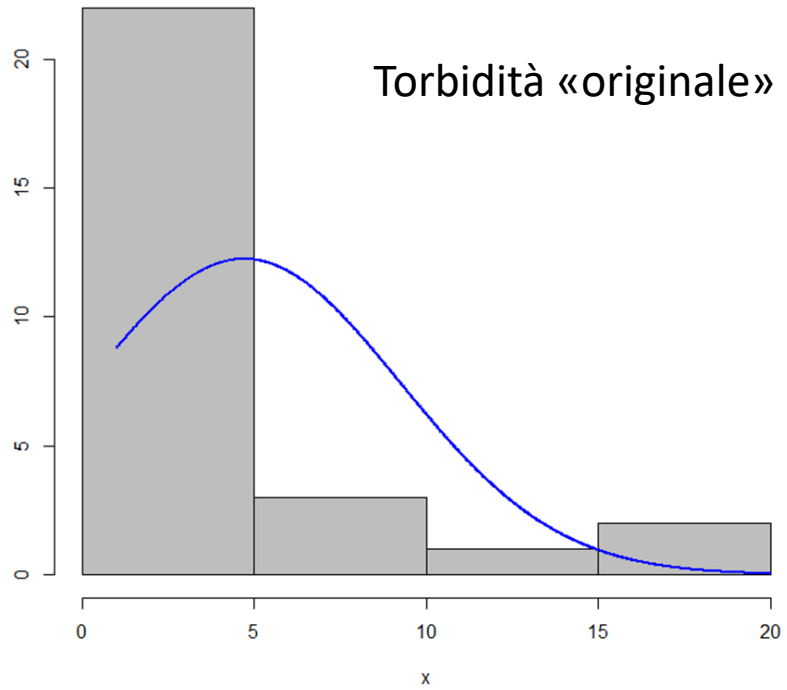
C'è un lieve miglioramento,
ma non soddisfacente

	Media	Dev Std	Mediana	Q1	Q3	Min	Max	Asimmetria	Curtosi
Turbidity	4.7	4.56	4	1.5	5	1	20	1.81	2.87
Sqrt(Turbidity)	1.98	0.9	2	1.22	2.24	1	4.47	1.06	0.49

Può essere calcolata su dati >= 0.



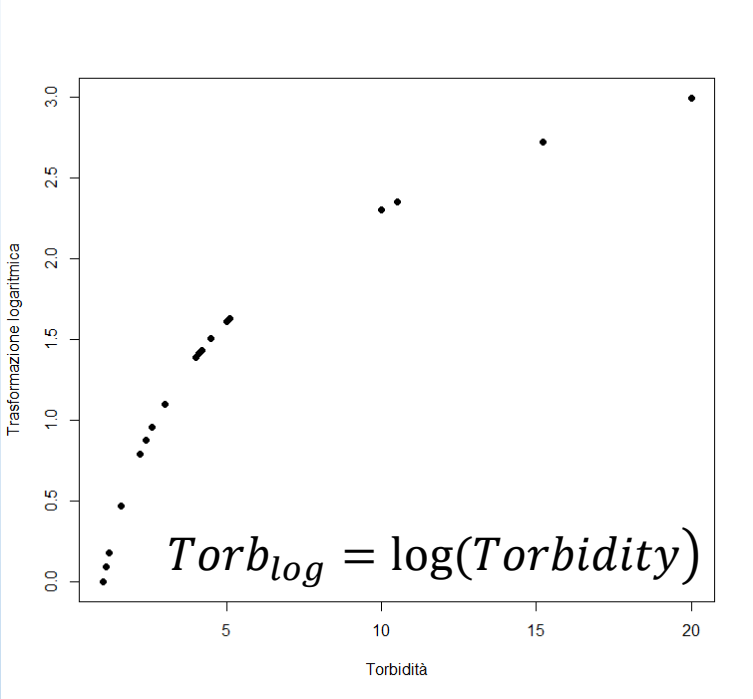


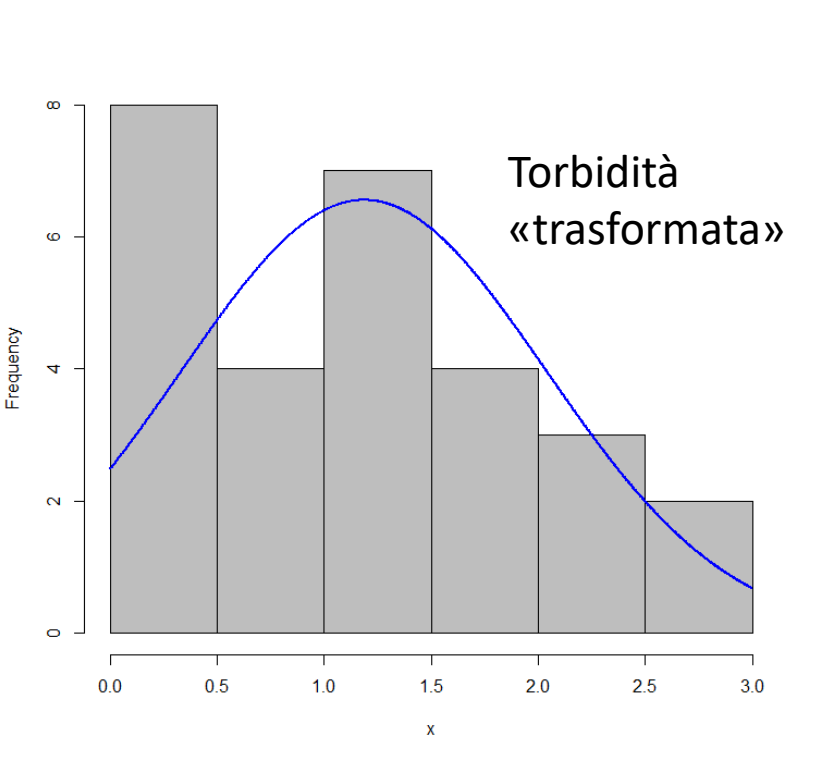


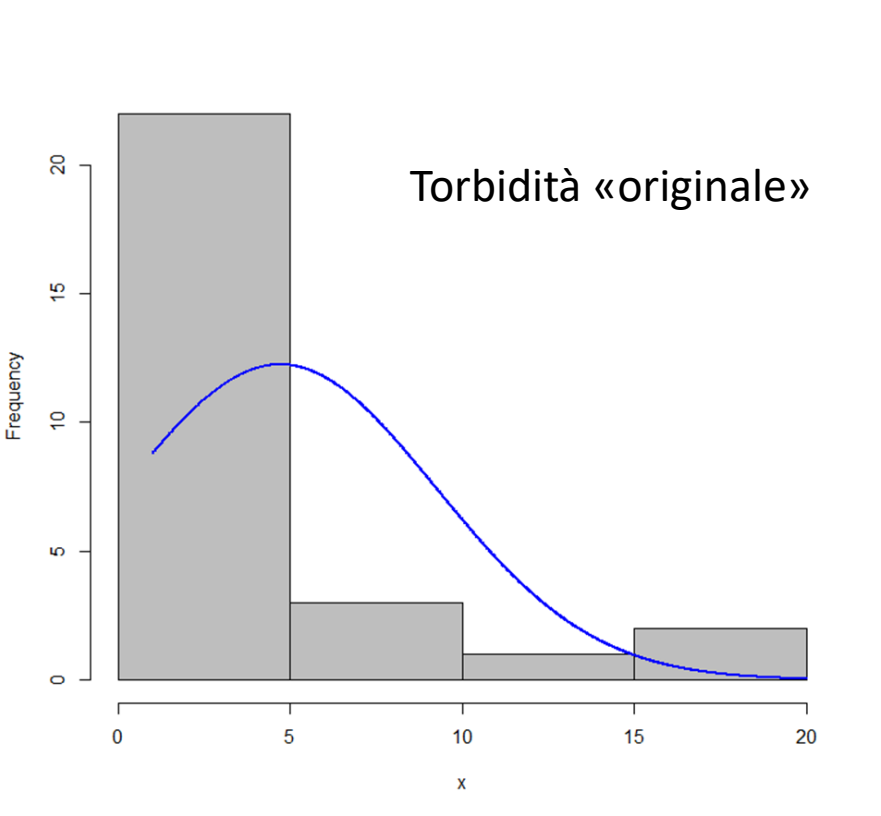
C'è un discreto miglioramento,
ma ancora non del tutto soddisfacente

	Media	Dev Std	Mediana	Q1	Q3	Min	Max	Asimmetria	Curtosi
Turbidity	4.7	4.56	4	1.5	5	1	20	**1.81**	**2.87**
Radcub(Turbidity)	1.54	0.46	1.59	1.14	1.71	1	2.71	**0.79**	**-0.10**

Può essere calcolata anche su dati negativi e che contengono il valore zero.







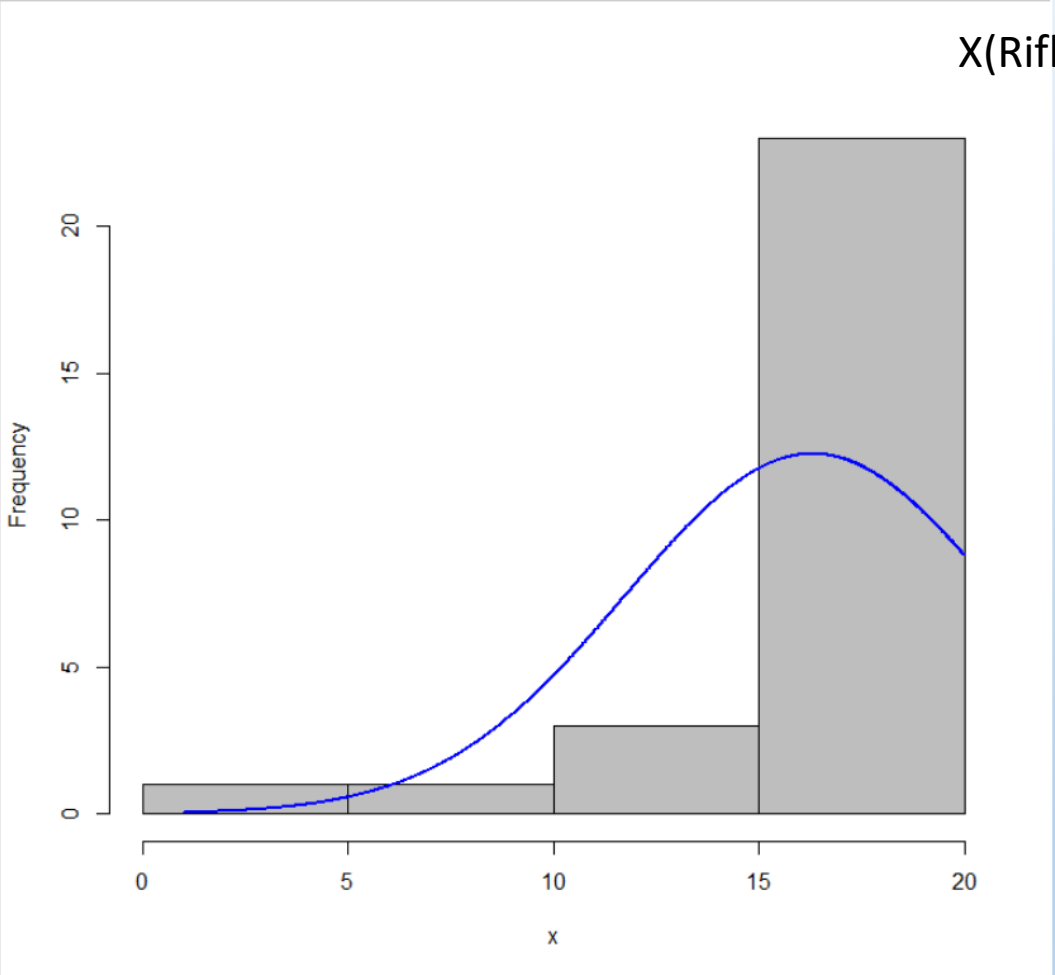
Decisamente meglio: visto il tipo di dati è probabilmente il risultato migliore che possiamo ottenere.

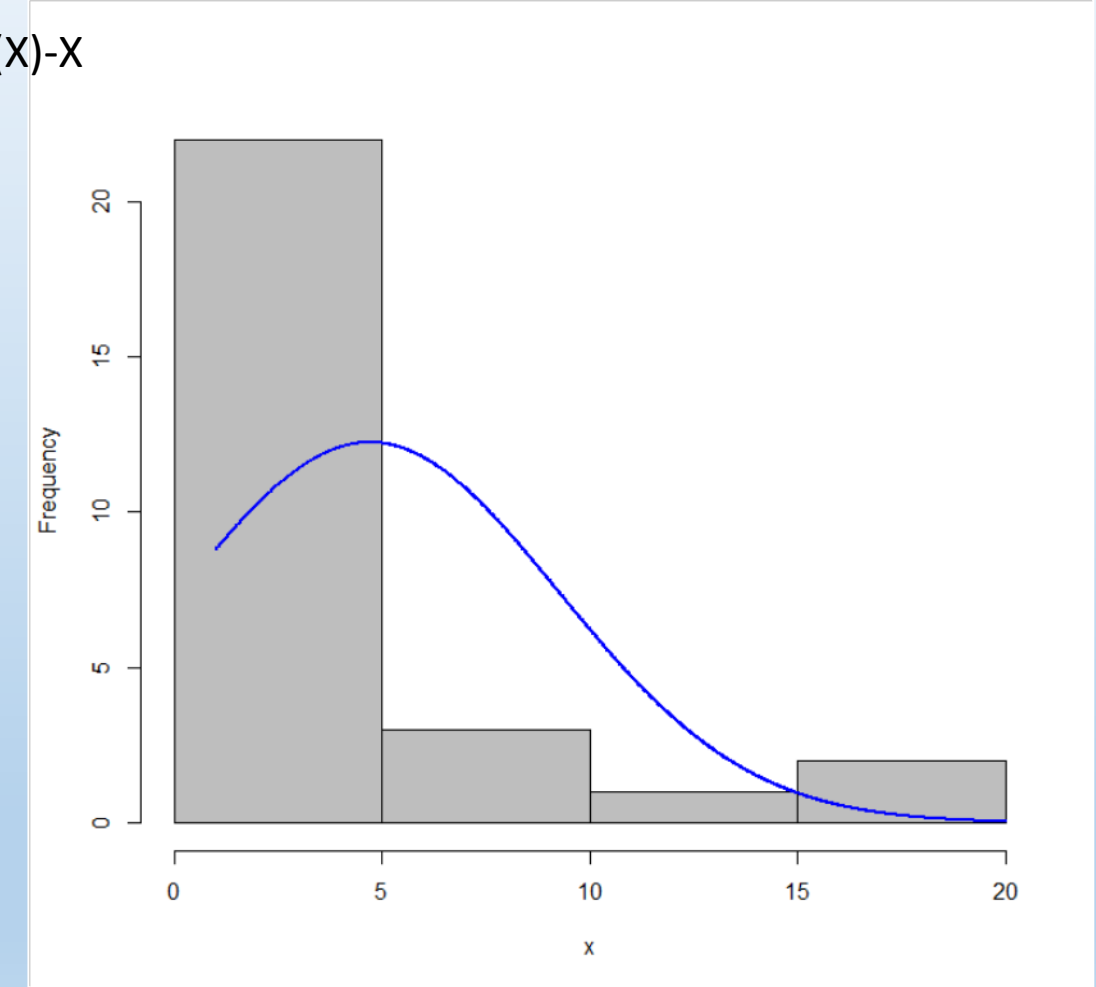
	Media	Dev Std	Mediana	Q1	Q3	Min	Max	Asimmetria	Curtosi
Turbidity	4.7	4.56	4	1.5	5	1	20	**1.81**	**2.87**
log(Turbidity)	1.19	0.85	1.39	0.4	1.61	0	3	**0.28**	**-0.87**

Logaritmo naturale: base è il numero di Nepero (circa 2.7); solo per numeri >0, se la distribuzione contiene 0 va aggiunta una costante

Teniamo presente che per le distribuzioni asimmetriche a sinistra, è opportuno prima di tutto creare la variabile «riflessa» e poi si possono applicare le trasformazioni viste (radice quadrata, cubica o logaritmo).

X(Riflessa)=1+Max(X)-X





Ci poi sono alcune funzioni che possiamo applicare sui dati per ottenere «*la migliore trasformazione possibile*», ma non entriamo in questo dettaglio tecnico.